Supplementary Information

Interpreting cancer genomes using systematic host perturbations by tumour virus proteins

Orit Rozenblatt-Rosen[1,2*#], Rahul C. Deo[3,4*#], Megha Padi[5,6*#], Guillaume Adelmant[3,7*#], Michael A. Calderwood[8,9,10*^], Thomas Rolland[8,9*^], Miranda Grace[11*^], Amélie Dricot[8,9*^], Manor Askenazi[3,7*^], Maria Tavares[1,2,7*^], Sam Pevzner[8,9,12^], Fieda Abderazzaq[5*], Danielle Byrdsong[8,9*], Anne-Ruxandra Carvunis[8,9], Alyce A. Chen[11*], Jingwei Cheng[1,2], Mick Correll[5], Melissa Duarte[8,10*], Changyu Fan[8,9*], Mariet C. Feltkamp[13], Scott B. Ficarro[3,7*], Rachel Franchi[8,14*], Brijesh K. Garg[3,7*], Natali Gulbahce[8,15,16*], Tong Hao[8,9*], Amy M. Holthaus[10*], Robert James[8,9*], Anna Korkhin[1,2*], Larisa Litovchick[1,2*], Jessica C. Mar[5,6*], Theodore R. Pak[17], Sabrina Rabello[2,8,15*], Renee Rubio[5*], Yun Shen[8,9*], Saurav Singh[3,7], Jennifer M. Spangle[11*], Murat Tasan[3,17*], Shelly Wanamaker[8,9,14], James T. Webber[3,7], Jennifer Roecklein-Canfield[8,14], Eric Johannsen[10*], Albert-László Barabási[2,8,15*], Rameen Beroukhim[2,18,19], Elliott Kieff[10*], Michael E. Cusick[8,9*], David E. Hill[8,9*], Karl Münger[11*], Jarrod A. Marto[3,7*], John Quackenbush[5,6*], Frederick P. Roth[3,8,17*], James A. DeCaprio[1,2*], Marc Vidal[8,9*]

*Member of the Genomic Variation and Network Perturbation Center of Excellence in Genomic Science, Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA.

[#]These authors contributed equally to this work and should be considered co-first authors.
[^]These authors contributed equally to this work and should be considered co-second authors.
Additional co-authors are listed alphabetically.

Correspondence and request for materials should be addressed to:
D.E.H. (david_hill@dfci.harvard.edu), K.M. (kmunger@rics.bwh.harvard.edu),
J.A.M. (jarrod_marto@dfci.harvard.edu), J.Q. (johnq@jimmy.harvard.edu),
F.P.R. (fritz.roth@utoronto.ca), J.A.D. (james_decaprio@dfci.harvard.edu),
M.V. (marc_vidal@dfci.harvard.edu).

[1]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA.

[2]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA.

[3]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA.

[4]Cardiovascular Research Institute, Department of Medicine and Institute for Human Genetics, University of California, San Francisco, California 94143, USA.

[5]Center for Cancer Computational Biology (CCCB), Department of Biostatistics and Computational Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA.

[6]Department of Cancer Biology, Dana-Farber Cancer Institute and Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA.

[7]Blais Proteomics Center and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA.

[8]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA.

[9]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.

[10]Infectious Diseases Division, The Channing Laboratory, Brigham and Women's Hospital and Departments of Medicine and of Microbiology and Immunobiology, Harvard Medical School, Boston, Massachusetts 02115, USA.

[11]Division of Infectious Diseases, Brigham and Women's Hospital and Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA.

[12]Biomedical Engineering Department, Boston University and Boston University School of Medicine, Boston, Massachusetts 02118, USA.

[13]Department of Medical Microbiology, Leiden University Medical center, Leiden, The Netherlands.

[14]Department of Chemistry, Simmons College, Boston, Massachusetts 02115, USA.

[15]Center for Complex Networks Research (CCNR) and Department of Physics, Northeastern University, Boston, Massachusetts 02115, USA.

[16]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, California 94158, USA.

[17]Donnelly Centre, University of Toronto and Lunenfeld Research Institute, Mt. Sinai Hospital, Toronto, M5G 1X5 Ontario, Canada.

[18]Department of Medical Oncology and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA.

[19]The Broad Institute, Cambridge, Massachusetts 02142, USA.

Contents

Supplementary Figures and Legends

Supplementary Methods

I. Pathway enrichment analysis
J. Microarray preprocessing and differential expression
K. Microarray gene and sample clustering
L. Predicting cell-specific transcription factor binding sites
M. Transcription factor binding site (TFBS) enrichment analysis
N. Randomization of gene clusters for enrichment analyses
O. Evaluating predictive power of regulatory cascades
P. Building a probabilistic map of IMR-90 specific RBPJ binding sites
Q. Notch pathway enrichment analysis
R. Identification of loci implicated in familial and somatic cancer
S. Testing enrichment of gene sets for cancer genes
T. Viral target overlap with candidate cancer genes identified by transposon screens
U. Comparison of viral interactome and prioritisation of cancer genes

Supplementary Tables (available on Nature website)

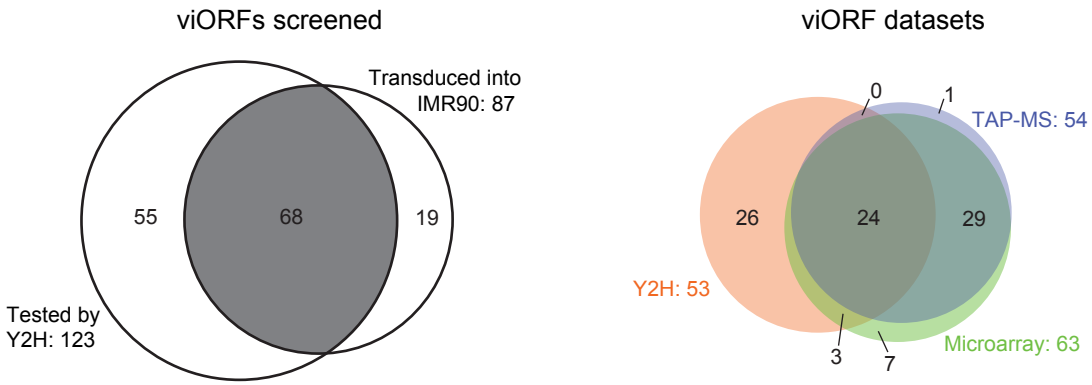
1	List of viral ORFs	Summary of all viral ORFs tested, including all information for each viORF and the data generated for each ORF.
2	Y2H dataset	Summary of virus-host binary interactions identified by Y2H screens.
3	Host proteins targeted by viral proteins in Y2H	List of host proteins in the Y2H network that are significantly targeted or untargeted by viral proteins.
4	TAP dataset	List of virus-host protein associations identified in two independent TAP-MS experiments.
5	GO cluster summary	List of the GO terms that are enriched in each cluster of differentially regulated genes, and the adjusted P values and odds ratios associated with them.
6	KEGG cluster summary	List of all the KEGG terms that are enriched in each cluster of differentially regulated genes, and the adjusted P values and odds ratios associated with them.
7	TF list by Y2H,TAP,Array	Virally targeted TFs with an enriched number of predicted binding sites for genes within each cluster. Listed TFs come from either associations through TAP-MS, interactions through Y2H, or are differentially expressed in response to viral ORF expression.
8	COSMIC Classic annotation	COSMIC Classic genes are annotated based on literature review as tumour suppressor, oncogene or both. If no call could be made, the designation is left as unclear.
9	VirHost dataset	947 VirHost proteins identified through TAP-MS association (3 or more unique peptides), Y2H interaction, or as differentially expressed TFs from motif enrichment analysis.
10	Transposon Candidates	Overlap between Sleeping Beauty and Murine Leukemia Virus transposon-based screens and VirHost data set.

11	Polyphen score	Cumulative POLYPHEN2 score calculated for each gene with one or more somatic mutations from genome-wide tumour sequencing data.
12	GWAS–VirHost, SCNA-DEL–VirHost, SCNA-AMP–VirHost intersections	Genes at intersection of VirHost data set with genes identified through cancer GWAS or through cancer SCNA analysis (amp=amplification, del=deletion).
13	PCR Primers and viORFs sequence	PCR primers used to generate viORFs from Ad5, HPVs and PyVs and sequence of each viORF.
14	Autoactivators	List of viORFs that act as auto-activators in Y2H or wNAPPA protein interaction assays.
15	Tandome	List of host proteins identified in more than 1% of 108 control TAP experiments (Tandome).
16	Y2H PRS, TAP-MS PRS	List of Y2H and TAP-MS Positive Reference Sets (PRS).
17	Pathway enrichment analysis	Gene Ontology term enrichment for viral targets identified by TAP-MS for Ad5, EBV, HPV and PyV viruses; Odds ratio and resampling-adjusted P-values are provided.
18	Microarray batches	Lists of the individual microarrays in the gene expression dataset, and the batch in which each one was processed.
19	Differentially regulated genes	Annotation of the most frequently perturbed host genes, with the cluster membership, fold changes, and adjusted P values in each viORF-expressing cell line relative to control cell line.
20	GWAS loci	Mapping of loci previously identified through cancer GWAS to nearby genes.
21	SCNA loci	Loci previously identfied through SCNA analysis of cancers, with corresponding statistical evidence (q-value and residual q-value) and genes within each interval.
22	wNAPPA results	wNAPPA scores for tested viORF interactions.

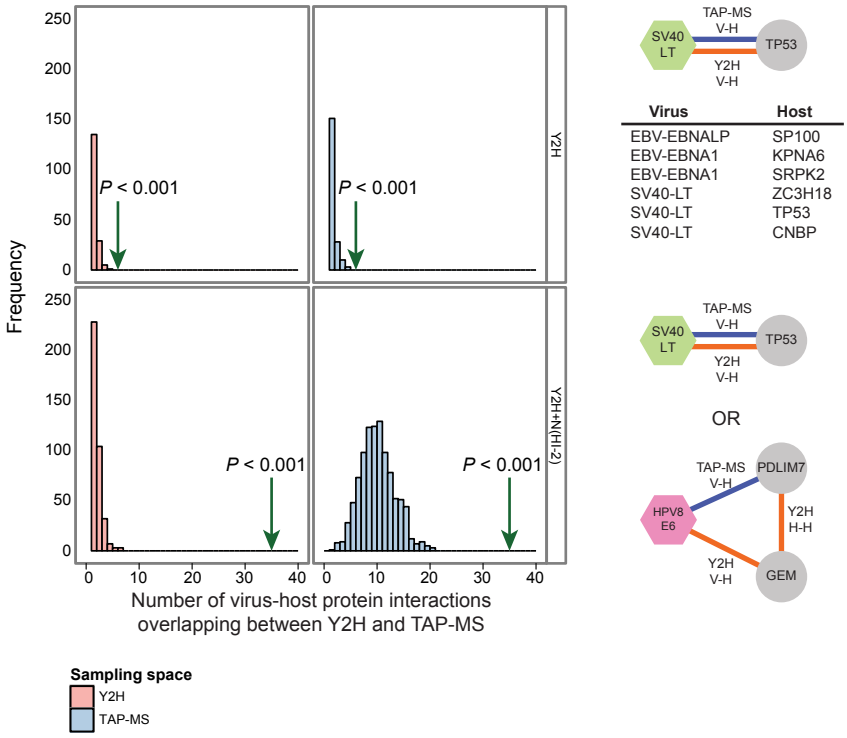
Supplementary Notes

1. Significantly targeted and untargeted proteins
2. HI-2 and analysis of overlaps between Y2H, TAP-MS and their respective PRS
3. Measurement of the precision of the Y2H dataset by wNAPPA
4. Network motif identification
5. Cellular growth phenotypes
6. Cancer pathways perturbed by viORFs

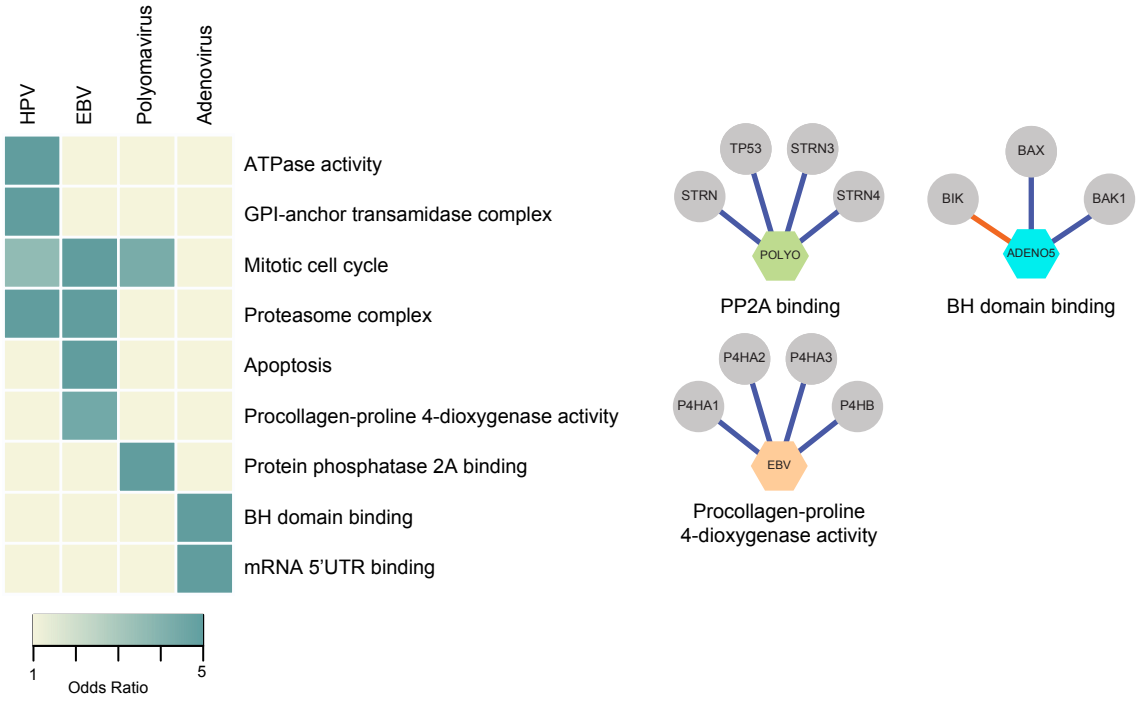
Supplementary References



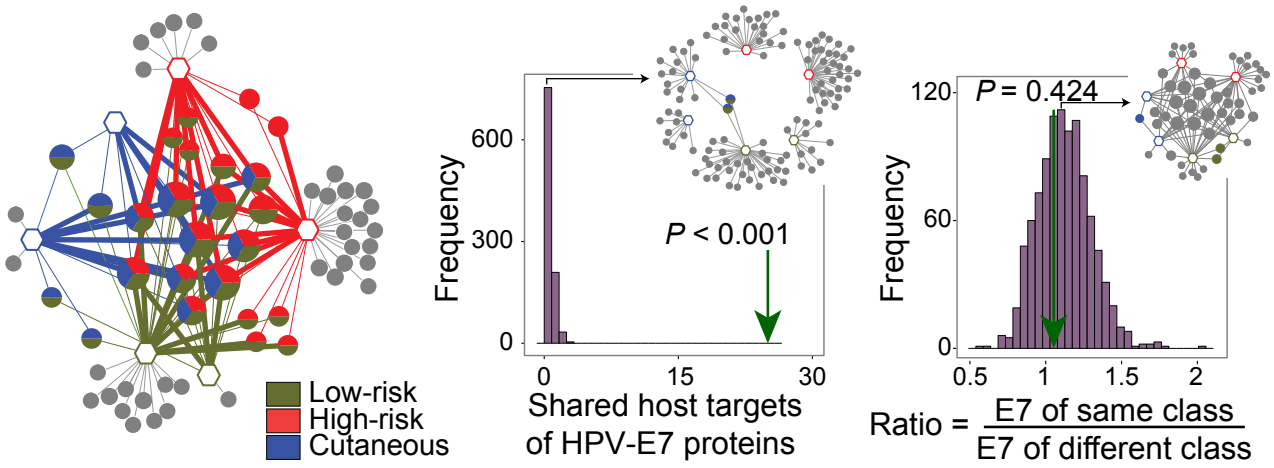
Supplementary Figure 1. Summary of viORFs screened. Overlap of the number of viORFs tested in the Y2H and cell line branches of the experimental pipeline (left Venn diagram), and the overlaps of viORFs that yielded data in each channel of the experimental pipeline (right Venn diagram).



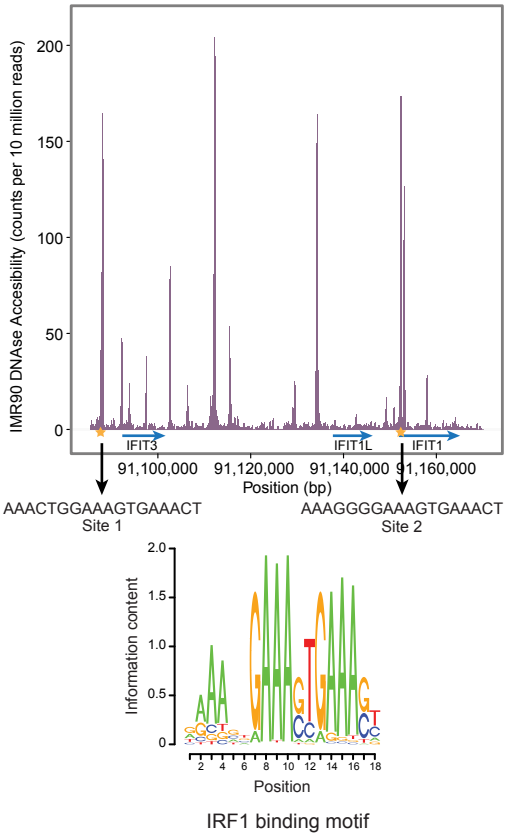
Supplementary Figure 2. Comparison of Y2H and TAP-MS datasets. Upper panels: number of virus-host interactions observed (green arrow) in Y2H and TAP-MS versus those seen by chance through random sampling of the Y2H (red) or TAP-MS (blue) search spaces, with six shared interactions observed listed. Lower panels: corresponding overlaps with expanded (Y2H+N(HI-2)) network, which includes human proteins “one hop” away in the HI-2 human-human interactome network.

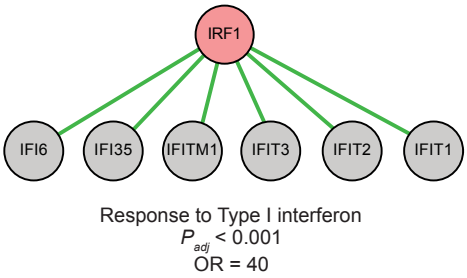


Supplementary Figure 3. Enrichment of GO terms for targeted host proteins. Enrichment of GO terms for host proteins physically interacting with viral proteins (Supplementary Table 17). Three examples are shown. All Odds Ratios higher than 5 were set to 5 for visualization purposes.

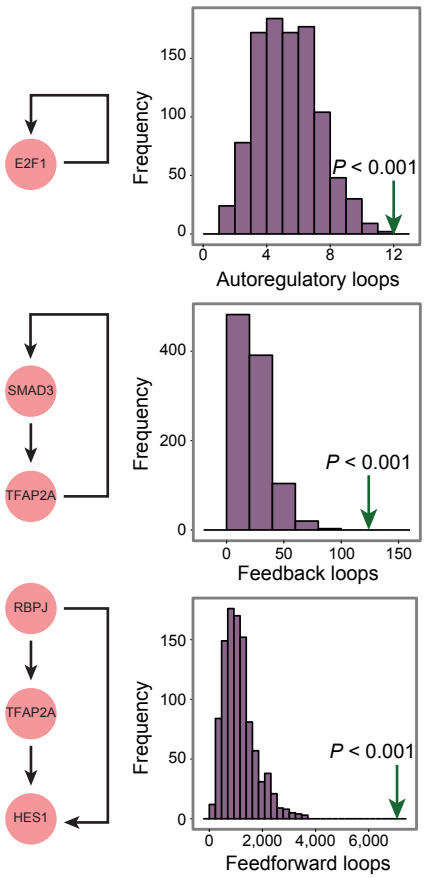


Supplementary Figure 4. Network of HPV E7 shared complexes. Network of co-complex associations of E7 viral proteins from six HPV types (hexagons, coloured according to disease class) with host proteins (circles). Host proteins that associate with two or more E7 proteins are coloured according to the disease class(es) of the corresponding HPV types. Circle size is proportional to the number of associations between host and viral proteins in the E7 networks. Viral-host protein co-complex associations (links) are weighted by the number of unique peptides detected for the host protein (thin links: 1-2; thick links: ≥ 3). Distribution plots of 1,000 randomised networks and experimentally observed data (green arrow) for the number of host proteins targeted at random by two or more proteins in the corresponding sub-networks (left histograms), or the ratio of the probability of a host protein being targeted by viral proteins from the same class to the probability it is targeted by viral proteins from different classes (right histograms). Representative random networks selected from these distributions are shown as insets in the histograms.

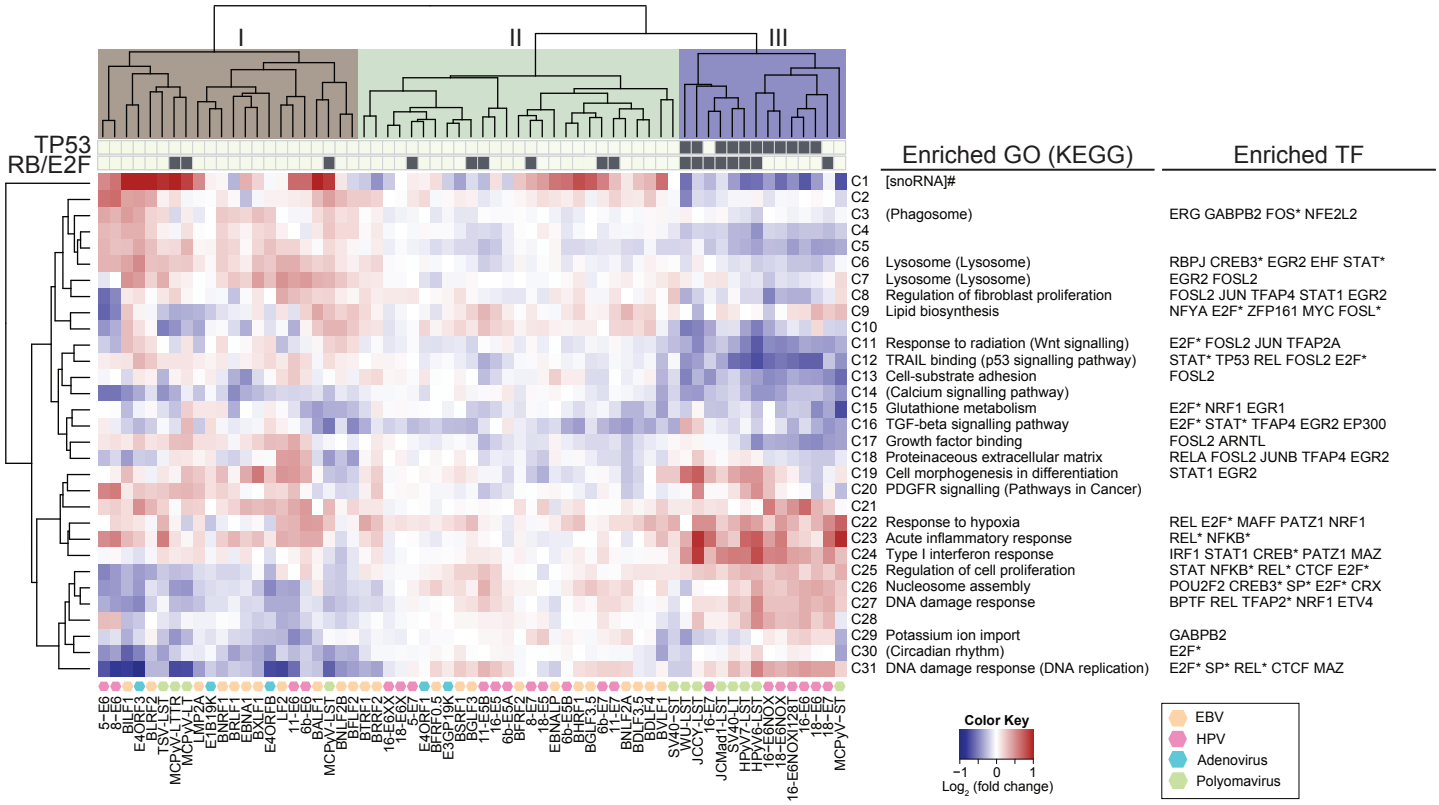




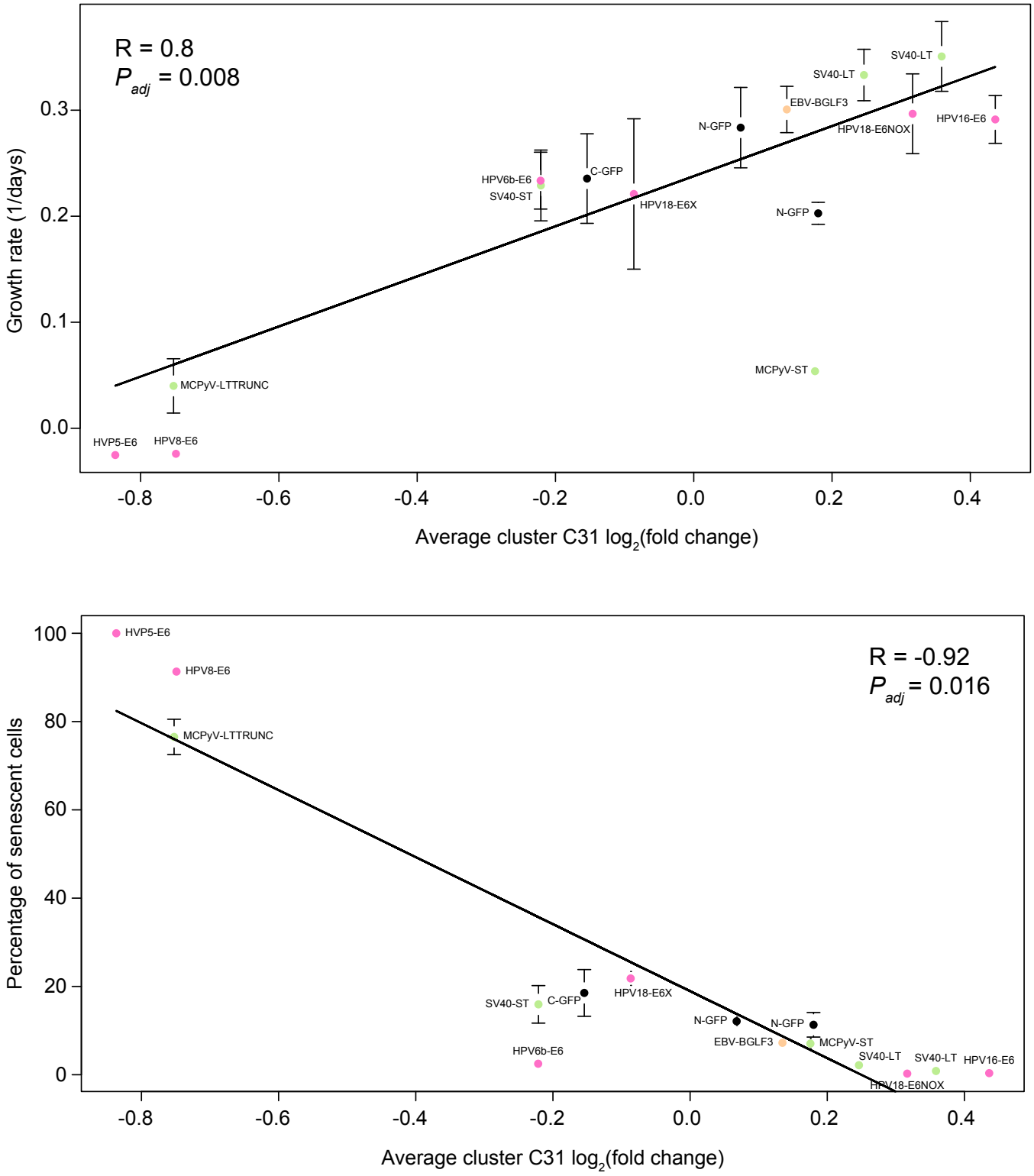
Supplementary Figure 5. Chromatin accessibility of IRF1 binding sites. DNase accessible canonical IRF1 binding sites in the promoters of interferon inducible genes highly expressed in response to Group III viral protein expression (left). Predicted targets of IRF1 within cluster C24 are significantly annotated for the GO term “Type I interferon response” (right).



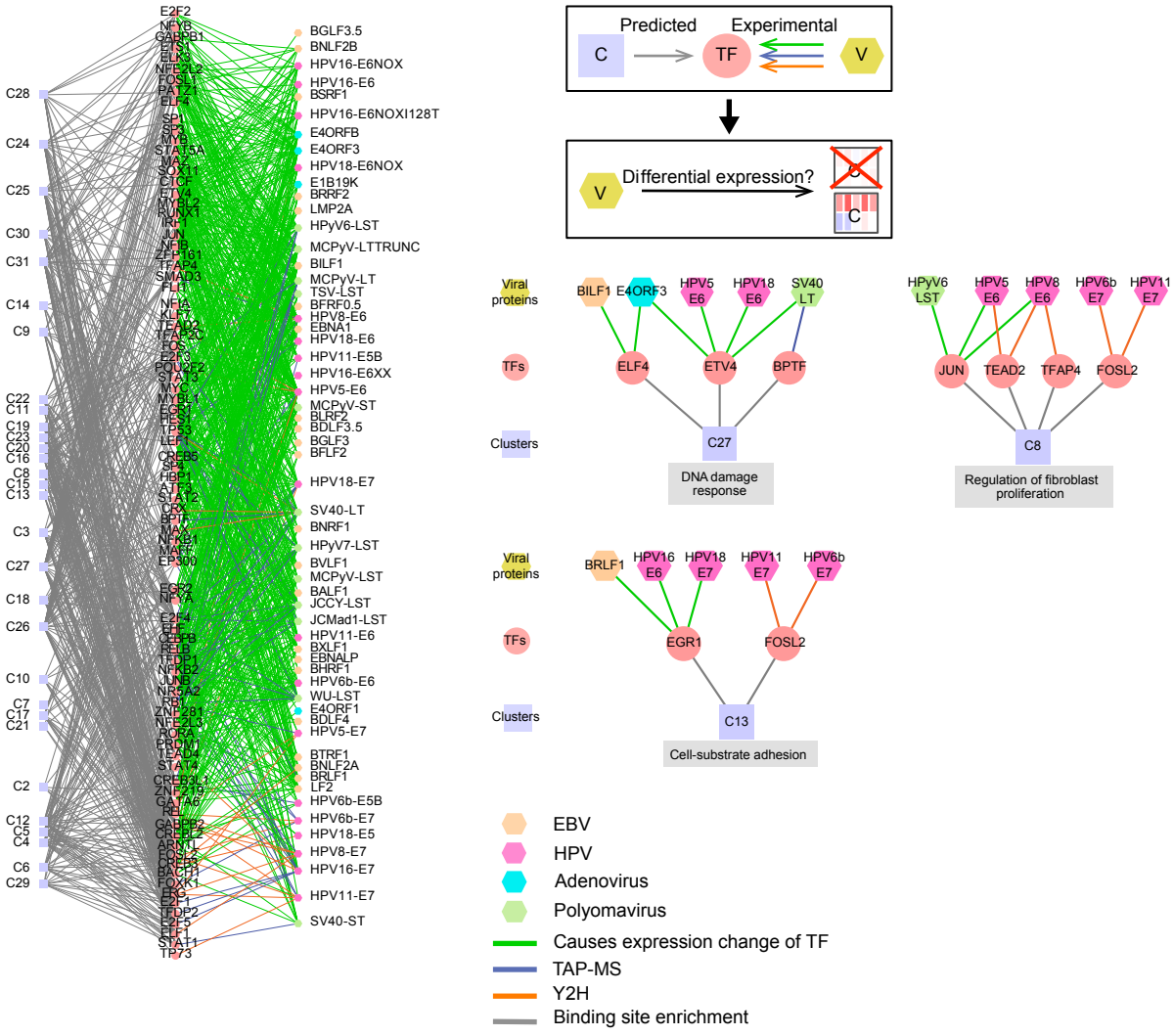
Supplementary Figure 6. Regulatory loops. Null distributions of autoregulatory (top), feedback (middle) and feedforward (bottom) motifs compared to the number of observed network motifs (green arrow) for enriched TFs.



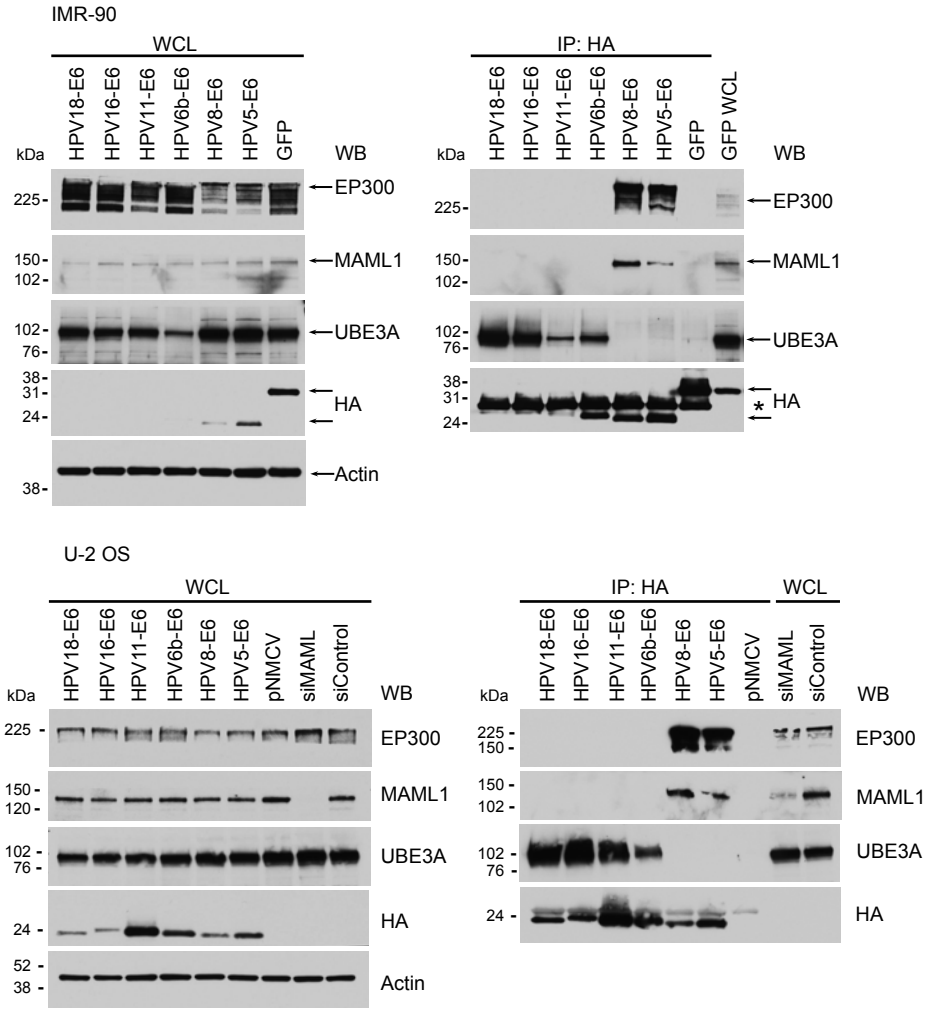
Supplementary Figure 7. Heatmap of transcriptome perturbations. Enlarged version of Fig. 2a. Enriched GO terms and KEGG pathways are listed adjacent to the numbered expression clusters. Transcription factors (TFs) with enriched binding sites and gene targets enriched for the listed GO and/or KEGG pathways that are physically associated with or differentially expressed in response to viral proteins are shown, with * denoting multiple members of a TF family. TFs are ranked by odds ratio (OR) for pathway enrichment, and up to 5 TFs are shown for any cluster. In cluster C1 eight of the nine transcripts are snoRNAs (# sign). Upper dendrogram is shaded by viORF grouping. Grey blocks show which viral proteins associate with the indicated host proteins.



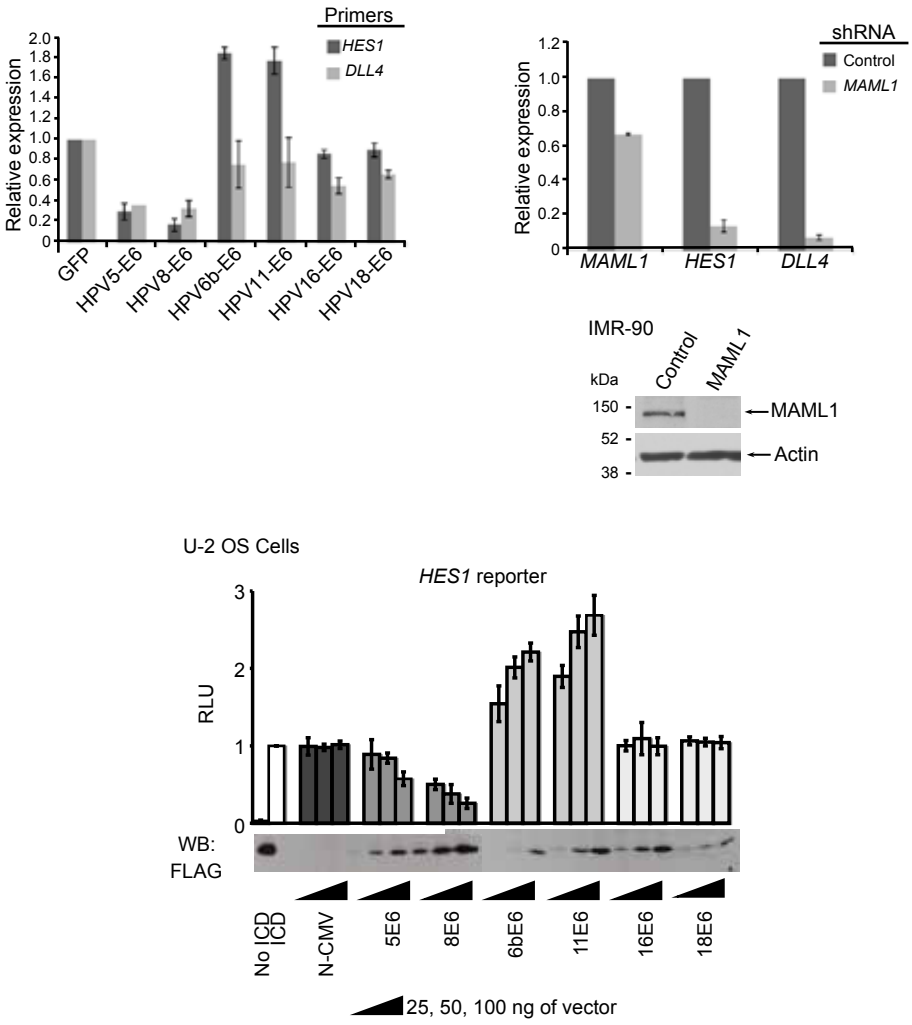
Supplementary Figure 8. Growth rate and senescence of selected IMR-90 cell lines expressing viORFs. Growth rate and senescence plotted against average expression of genes in cluster 31. Error bars in growth rate represent standard deviation across three samples in all cases except for MCPyV-LTTRUNC, which is across two samples, and HPV5-E6 and HPV8-E6, which represent one sample each. Error bars in senescence represent standard deviation across two samples.



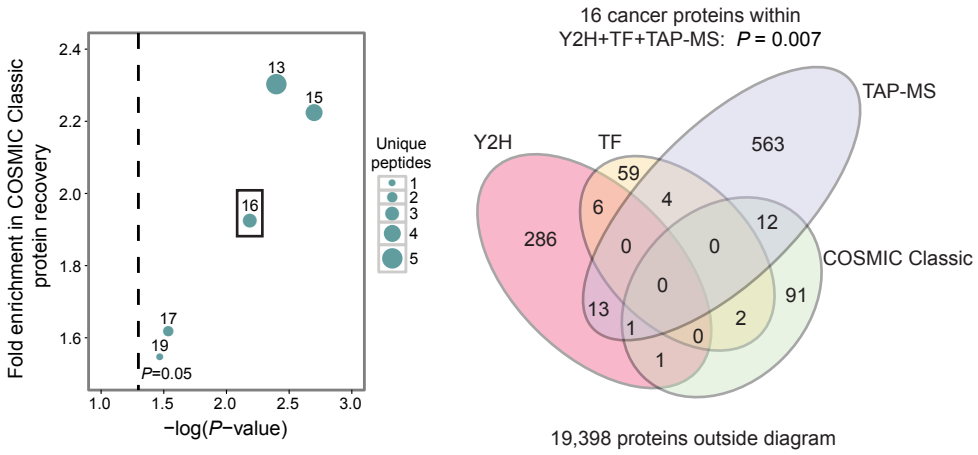
Supplementary Figure 9. Viral proteins to transcription factors to clusters. Network representation of all predicted viral protein-TF-cluster cascades (left). Schematic (right) shows how viral protein-TF-target gene network was constructed, with representative networks underneath.



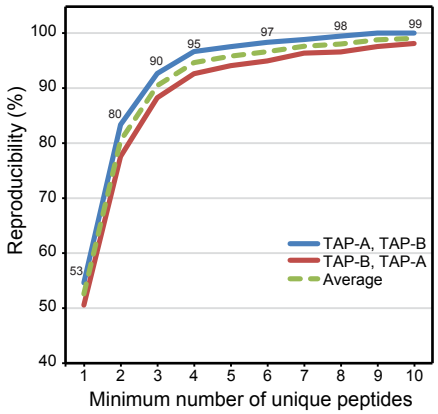
Supplementary Figure 10. MAML is targeted by E6 from HPV 5 and 8. Western blots of whole cell lysates (WCL) and co-immunoprecipitations of HPV E6 proteins in IMR-90 cells (upper panels) or U-2 OS cells (lower panels).



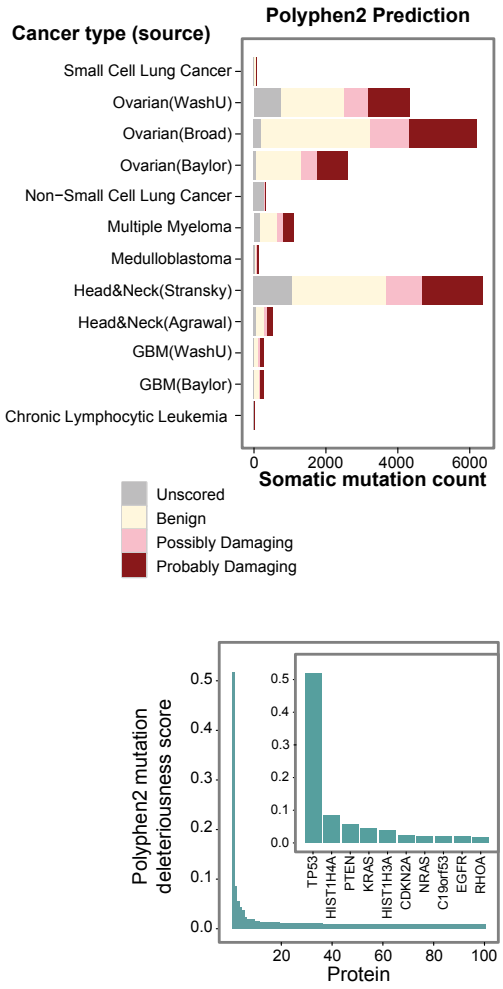
Supplementary Figure 11. The E6 protein from HPV 5 and 8 regulate *HES1* and *DLL4* expression. Upper panels: qPCR of Notch pathway responsive genes upon expression of E6 proteins from different HPV types (left panel) or upon knockdown of *MAML1* (right panel). Error bars represent standard deviation across two experiments and three experiments, respectively. Western blot of MAML1 in IMR-90 cells upon shRNA knockdown. Lower panel: *HES1* luciferase reporter assays of co-transfected Notch intracellular domain (Notch-ICD) and E6 proteins at increasing concentrations from different HPV types in U-2 OS cells. Error bars represent standard deviation across three experiments. Western blot (inset) depicts E6 protein expression from the indicated HPV type in a representative experiment.



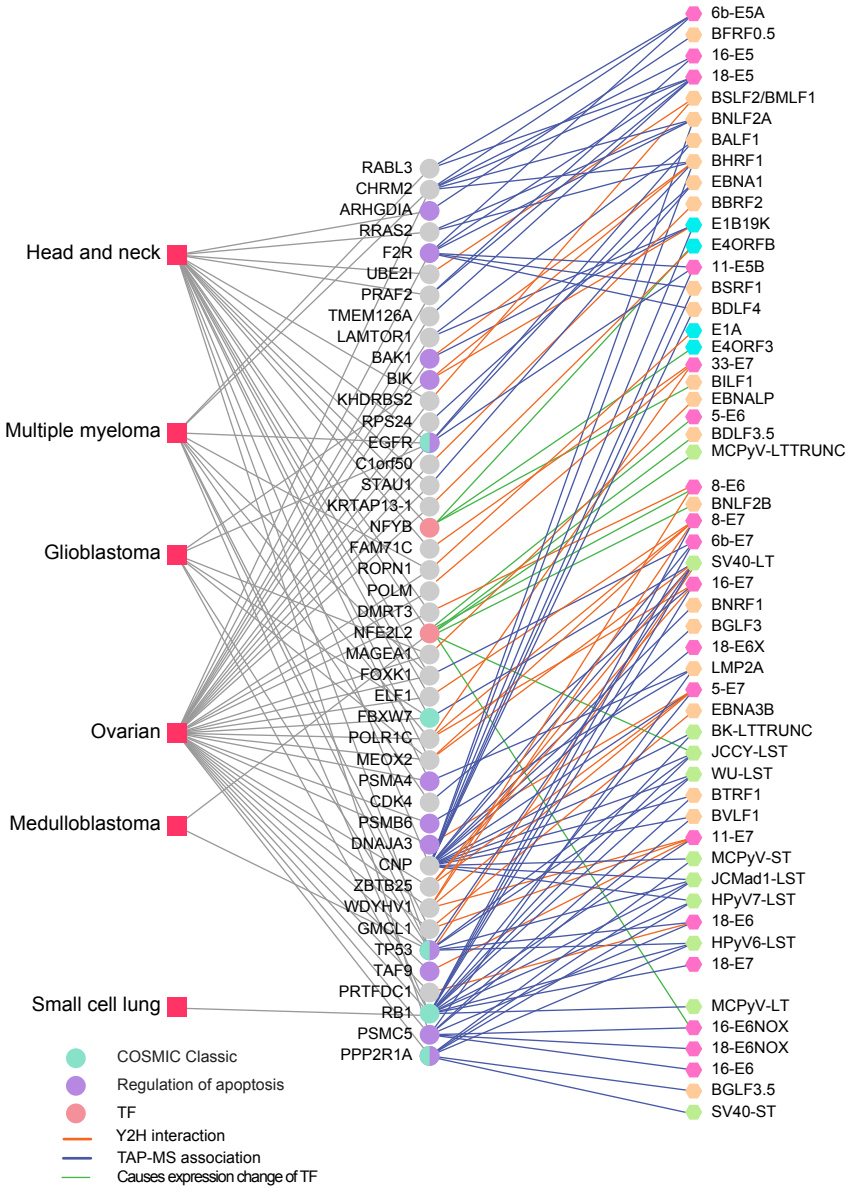
Supplementary Figure 12. Tumour viruses target cancer genes. Fold enrichment of COSMIC Classic genes in TAP-MS, Y2H and TF VirHost datasets as a function of the number of unique peptides (circle sizes) detected for host proteins identified through TAP-MS. Numbers of proteins in intersections of VirHost and COSMIC Classic genes are indicated. Venn diagram of overlaps between proteins identified by TAP-MS (≥3 unique peptides), Y2H, TF and COSMIC Classic genes.



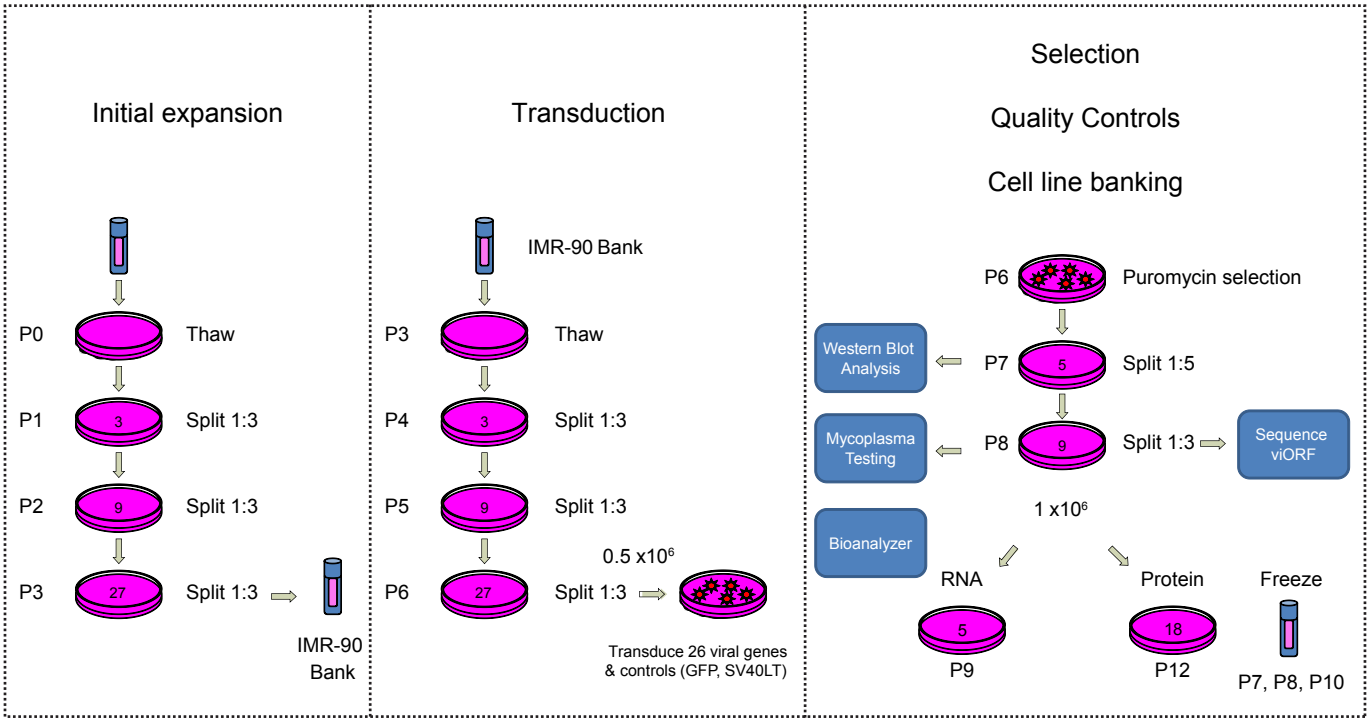
Supplementary Figure 13. Reproducibility of viral-host protein associations observed in replicate TAP-MS experiments as a function of the number of unique peptides detected. Data were plotted based on both experimental orientations (TAP-A then TAP-B, blue line and TAP-B then TAP-A, red line), with the average indicated by the green dashed line.



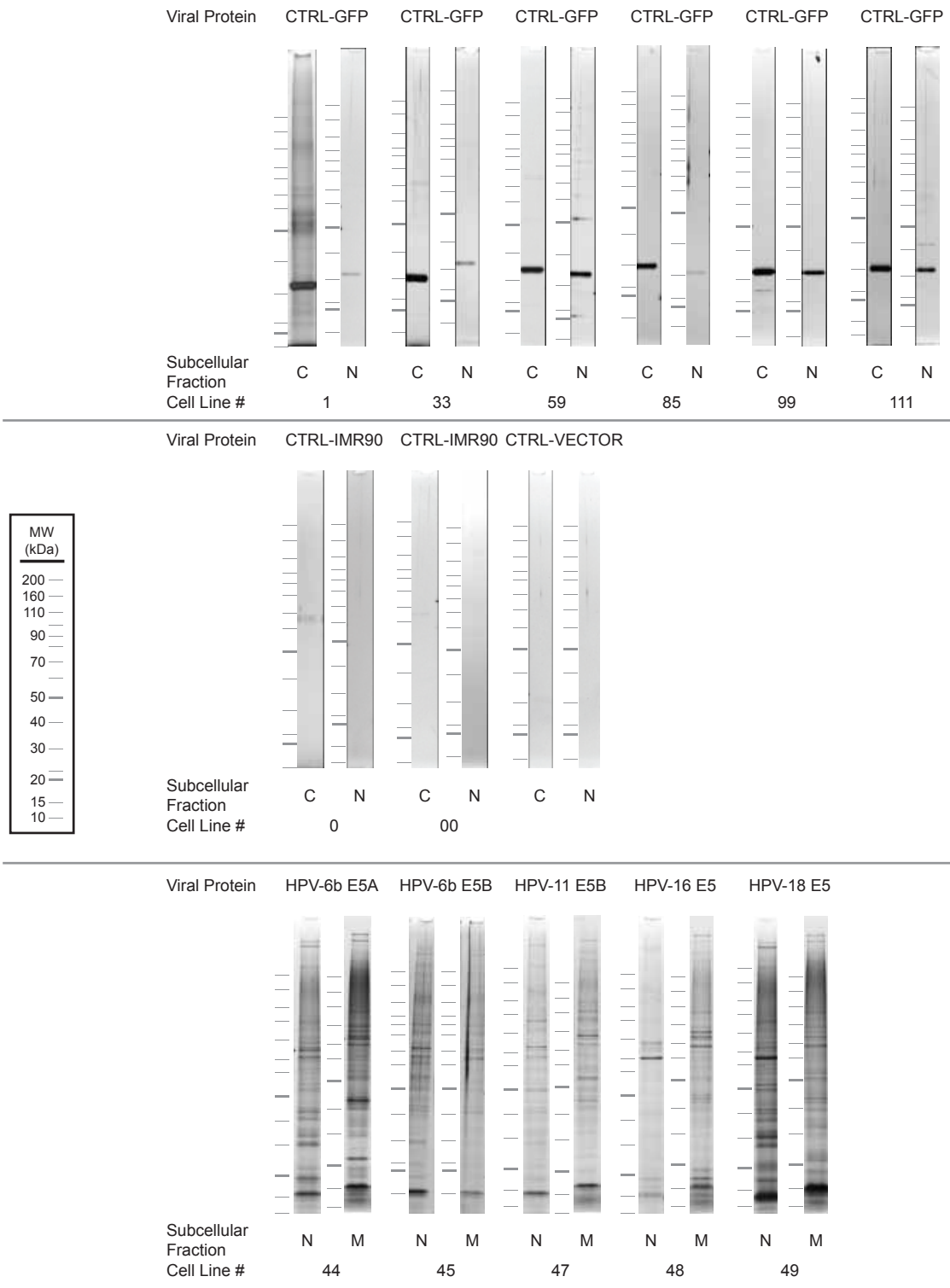
Supplementary Figure 14. Somatic mutations in tumour samples. Somatic mutations identified through genome-wide sequencing studies of human cancers and evaluated for impact using Polyphen2 (upper panel). Ranking of proteins with somatic mutations in cancer based on the cumulative inferred impact of all observed mutations (lower panel) (Supplementary Table 11). Inset: top ten ranked proteins.

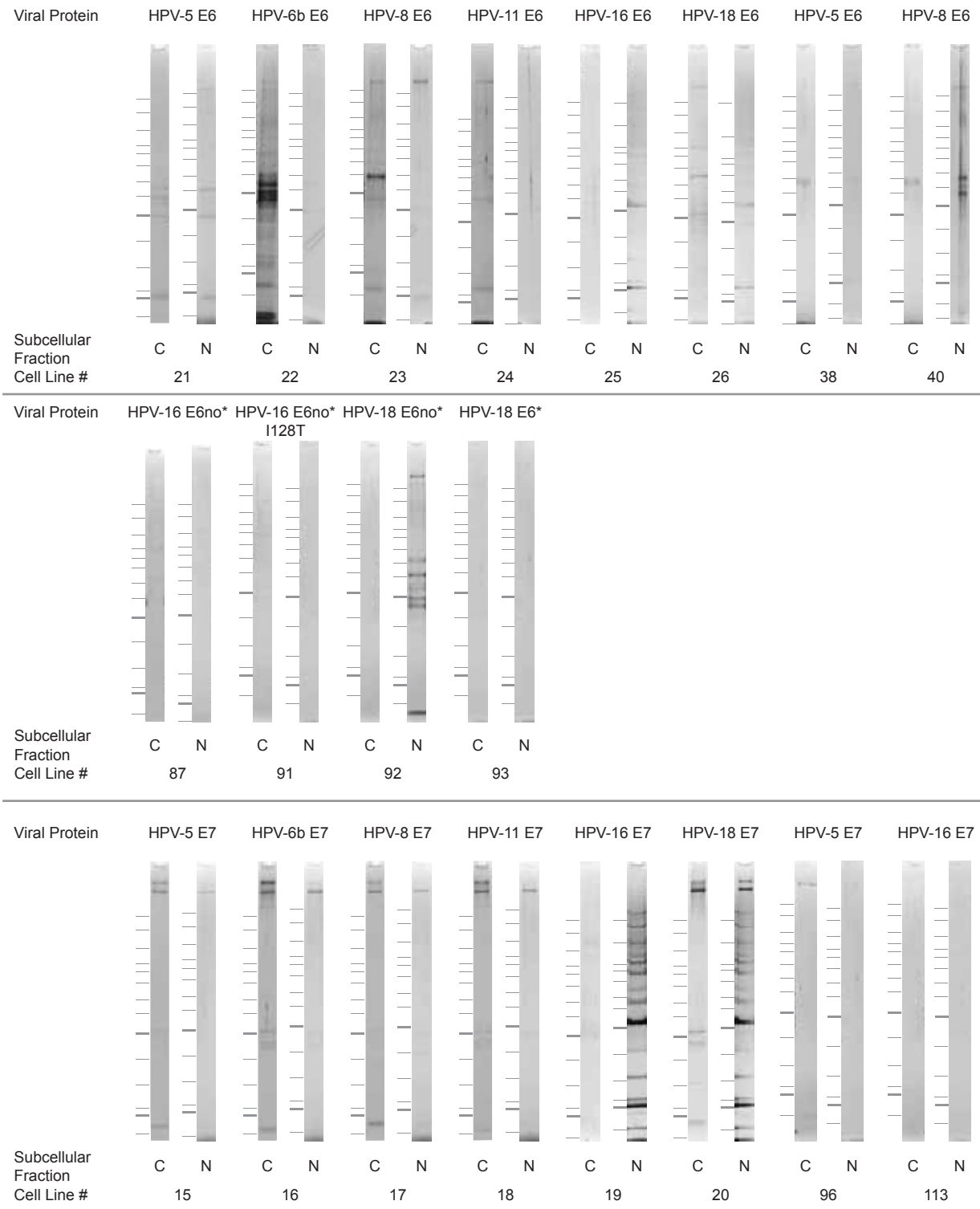


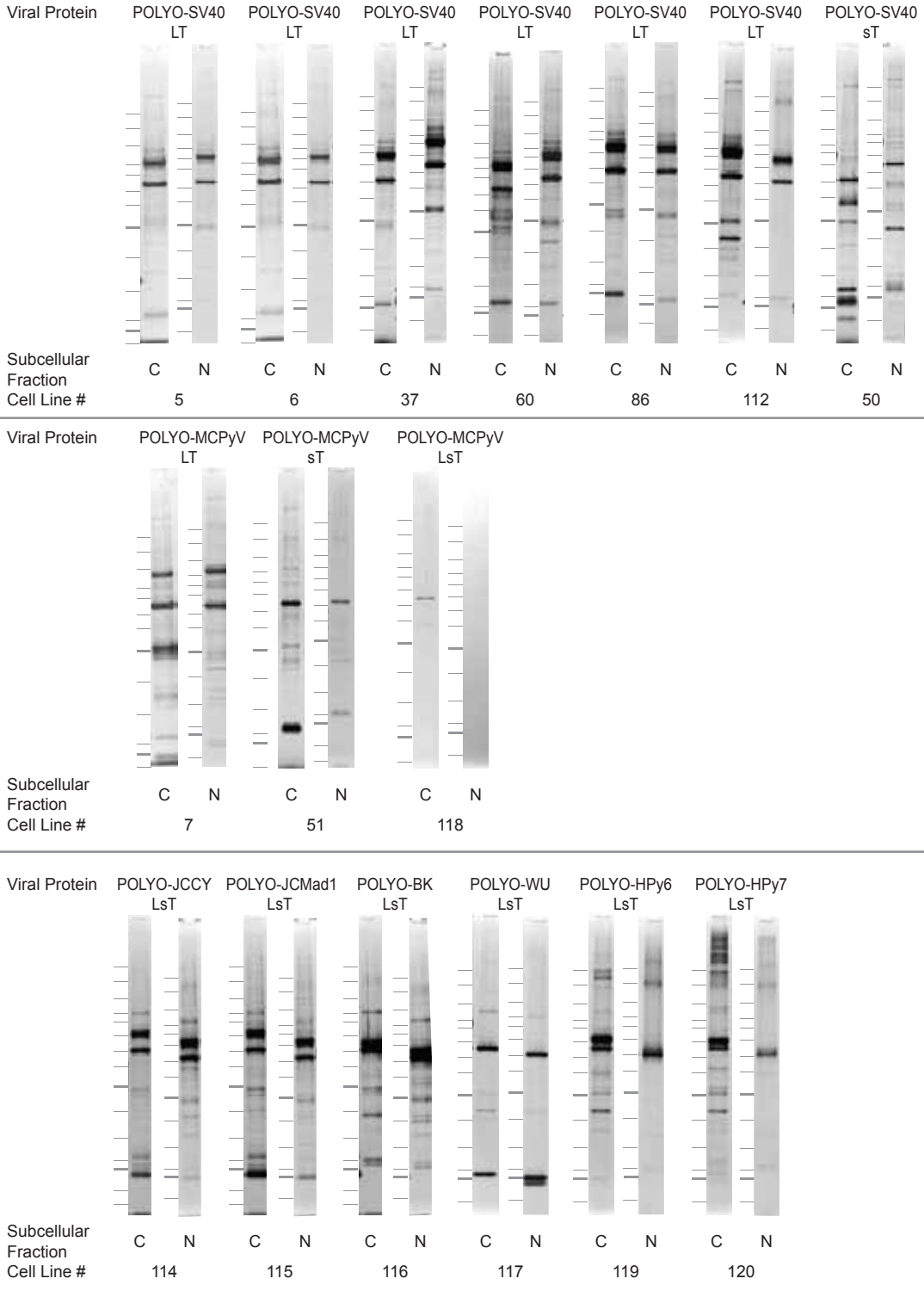
Supplementary Figure 15. Network of VirHostSM to host targets and cancers. Mapping of VirHostSM gene products to both tumours in which they are mutated (left) and to viral interactors (right). Proteins annotated with the GO term “regulation of apoptosis” indicated in purple.

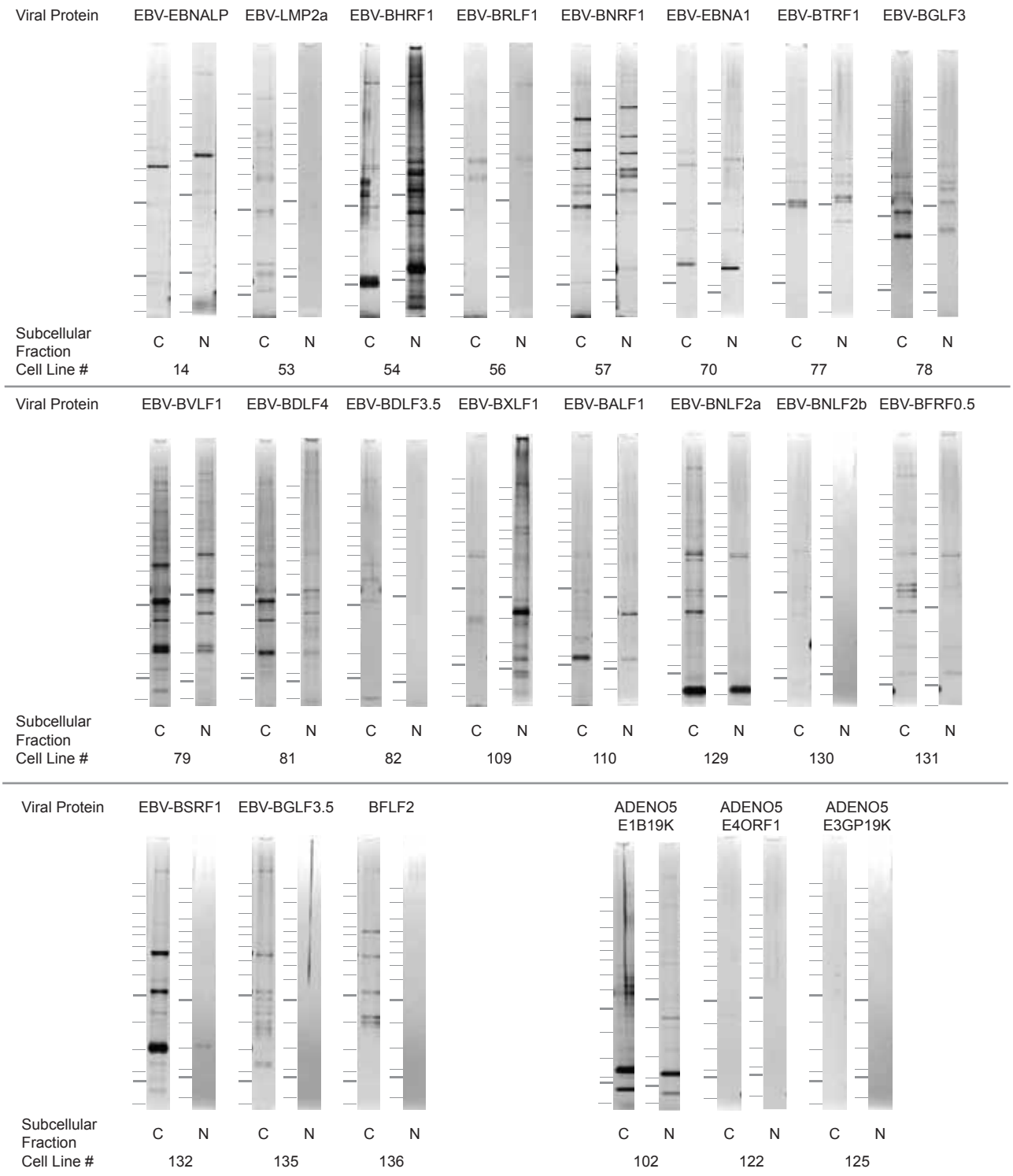


Supplementary Figure 16. The IMR-90 cell culture pipeline. Generation of the IMR-90 cell bank (left), and generation of IMR-90 cell lines expressing viORFs (middle) and quality control (QC) measures taken (right).

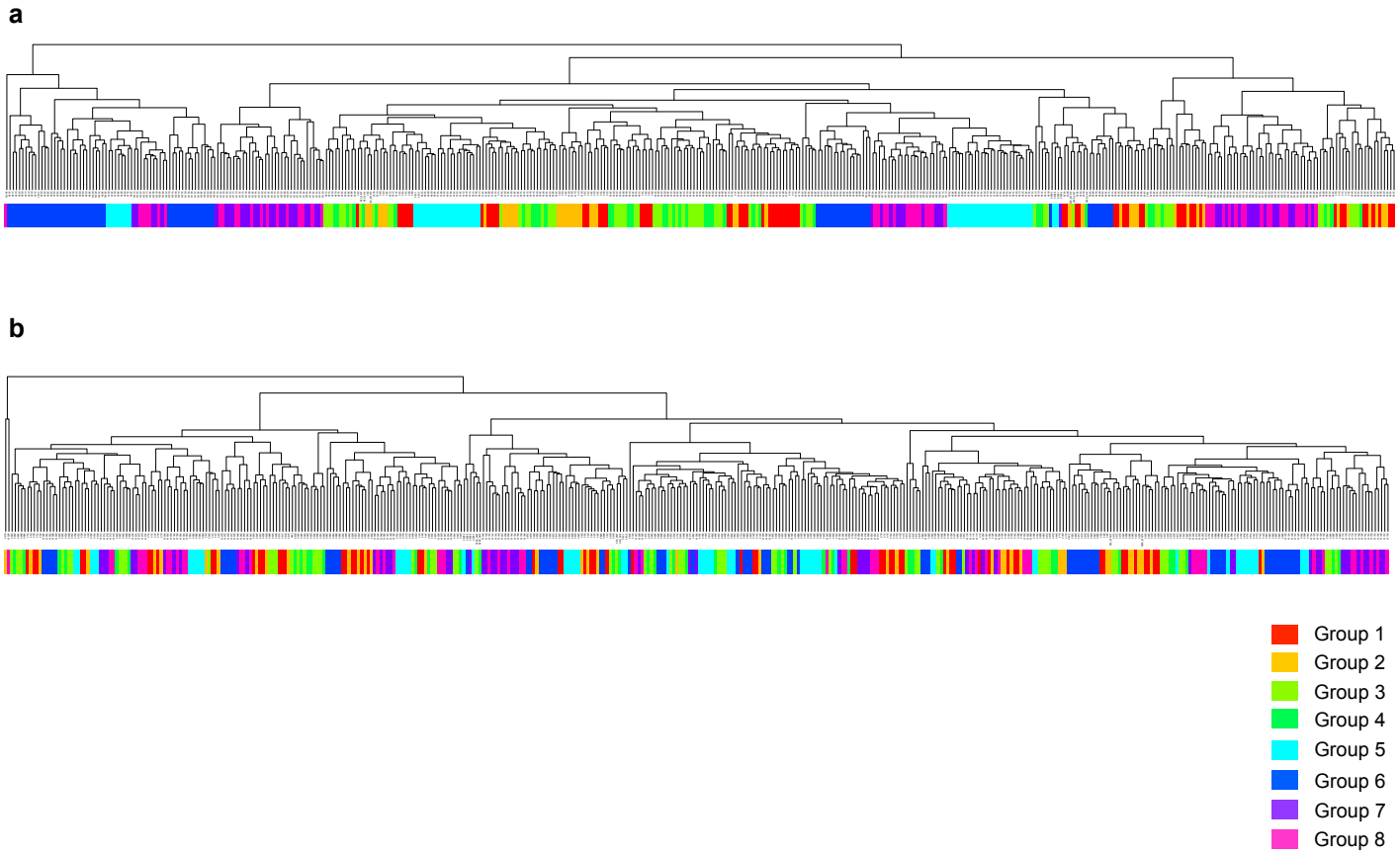

Viral Protein
CTRL-GFP
CTRL-GFP
CTRL-GFP
CTRL-GFP
CTRL-GFP
CTRL-GFP
Subcellular Fraction
C N C N C N C N C N C N
Cell Line #
1
33
59
85
99
111
Viral Protein
CTRL-IMR90
CTRL-IMR90
CTRL-VECTOR
MW (kDa)
200
160
110
90
70
50
40
30
20
15
10
Subcellular Fraction
C N C N C N
Cell Line #
0
00
Viral Protein
HPV-6b E5A
HPV-6b E5B
HPV-11 E5B
HPV-16 E5
HPV-18 E5
Subcellular Fraction
N M N M N M N M N M
Cell Line #
44
45
47
48
49


Viral Protein
HPV-5 E6
HPV-6b E6
HPV-8 E6
HPV-11 E6
HPV-16 E6
HPV-18 E6
HPV-5 E6
HPV-8 E6
Subcellular Fraction
C N C N C N C N C N C N C N C N
Cell Line #
21 22 23 24 25 26 38 40
Viral Protein
HPV-16 E6no*
HPV-16 E6no* I128T
HPV-18 E6no*
HPV-18 E6*
Subcellular Fraction
C N C N C N C N
Cell Line #
87 91 92 93
Viral Protein
HPV-5 E7
HPV-6b E7
HPV-8 E7
HPV-11 E7
HPV-16 E7
HPV-18 E7
HPV-5 E7
HPV-16 E7
Subcellular Fraction
C N C N C N C N C N C N C N C N
Cell Line #
15 16 17 18 19 20 96 113

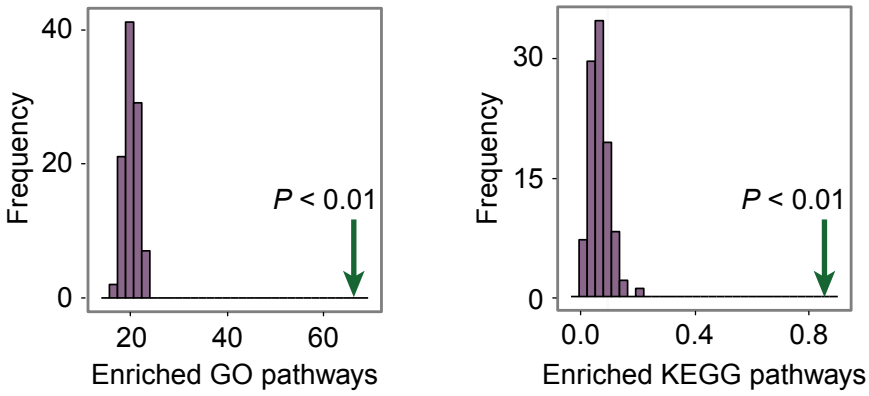

Viral Protein
POLYO-SV40 LT
POLYO-SV40 LT
POLYO-SV40 LT
POLYO-SV40 LT
POLYO-SV40 LT
POLYO-SV40 LT
POLYO-SV40 sT
Subcellular Fraction
C N C N C N C N C N C N C N
Cell Line #
5 6 37 60 86 112 50
Viral Protein
POLYO-MCPyV LT
POLYO-MCPyV sT
POLYO-MCPyV LsT
Subcellular Fraction
C N C N C N
Cell Line #
7 51 118
Viral Protein
POLYO-JCCY LsT
POLYO-JCMad1 LsT
POLYO-BK LsT
POLYO-WU LsT
POLYO-HPy6 LsT
POLYO-HPy7 LsT
Subcellular Fraction
C N C N C N C N C N C N
Cell Line #
114 115 116 117 119 120



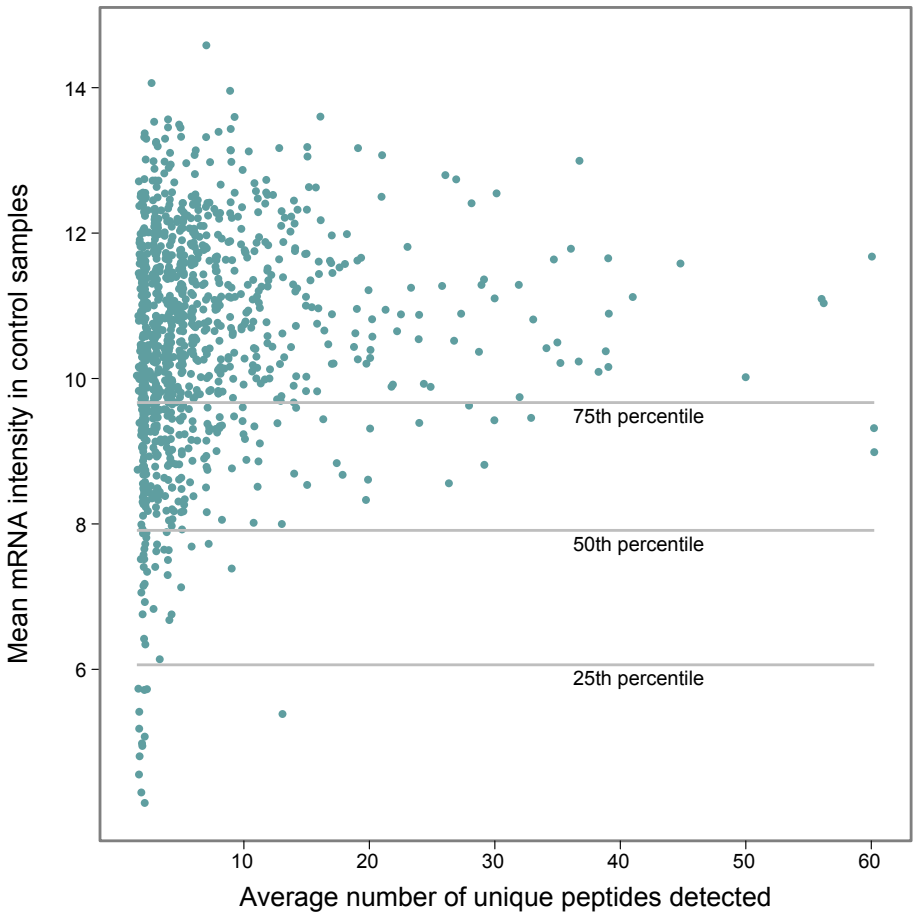
Supplementary Figure 17. Silver stain analyses of viral-host protein complexes. A 15% aliquot of the final eluate of each TAP was resolved on polyacrylamide gels then visualized by silver-stain (shown here for the first TAP replicate).



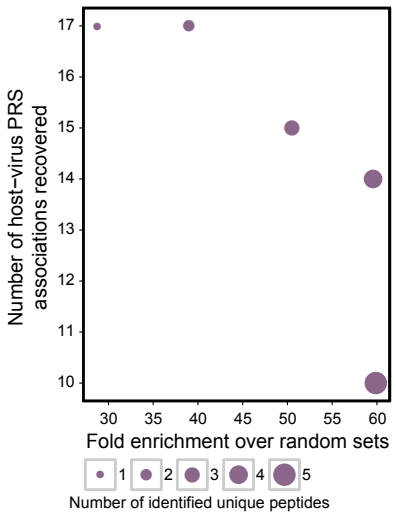
Supplementary Figure 18. Cluster propensity of microarrays before and after ComBat. Hierarchical clustering of all microarrays (a) before and (b) after applying ComBat, an algorithm used for removing batch effects in microarray data.



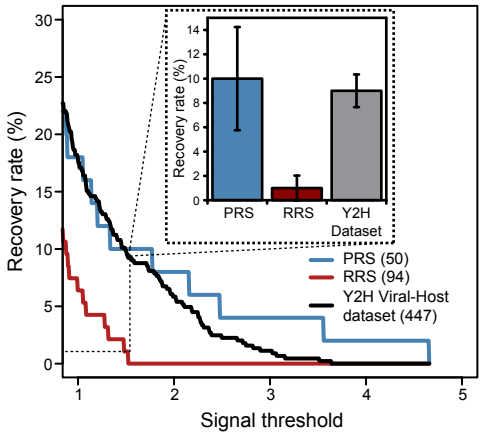
Supplementary Figure 19. Cluster coherence. Frequency of KEGG and GO pathway enrichment as compared to randomly generated clusters.



Supplementary Figure 20. mRNA expression bias. Viral targets identified through TAP-MS are biased towards highly expressed genes. Relationship between mRNA abundance in control samples of IMR-90 (y-axis) versus the number of unique peptides observed for viral associations for that protein (x-axis). Number of unique peptides is the average for two biological replicates. If more than one virus association is seen for a host protein, the association with the highest number of unique peptides is shown. Horizontal lines correspond to the indicated percentiles for mRNA abundance of all transcripts on the microarray.



Supplementary Figure 21. Overlap of viral-host protein pairs identified through TAP-MS with a literature-curated positive reference set. All points are significant at $P < 0.001$. Dot size reflects minimum number of unique peptide detections required for protein identification.



Supplementary Figure 22. Percentage of Y2H interacting protein pairs positive in wNAPPA assay at increasing assay signal for PRS, RRS and Y2H viral-host dataset. Inset: fraction of PRS or Y2H viral-host dataset positive in wNAPPA at a threshold of 1% RRS positive.

Supplementary Methods

A. Viral ORFeome cloning

Open-reading frame (ORF) clones encoding selected proteins from Adenovirus 5 (Ad5), seven human papillomaviruses, and nine polyomaviruses (Supplementary Table 1) were obtained by PCR-based Gateway recombinational cloning, following a protocol previously described for cloning Epstein-Barr Virus ORFs[4]. To generate viral entry clones ORFs were PCR-amplified with KOD HotStart Polymerase (Novagen). The PCR primers used contained attB1.1 and attB2.1 recombination sites fused to ~20 nucleotides of ORF-specific forward and reverse primers, respectively. Primers used to clone viORFs are listed (Supplementary Table 13). PCR products were then transferred into pDONR223 by a Gateway BP reaction, followed by transformation into chemically competent *E. coli* DH5α cells, selecting for spectinomycin resistance[30].

Sequence-verified entry clone viral ORFs (viORFs) were transferred by Gateway LR recombinational cloning (Invitrogen) into appropriate expression vectors. Recombination products were directly transformed into *E. coli* (DH5α-T1^R strain) via selection for ampicillin resistance in liquid LB media. Plasmid DNA was extracted from *E. coli* grown overnight in a 96-well format using a Qiagen BioRobot 8000. The prepared plasmid DNA was used for transformations into yeast cells or for generation of retrovirus for subsequent transductions into IMR-90 human cells.

For assay of protein interactions by yeast two-hybrid (Y2H), viORFs were introduced into both pDEST-DB and pDEST-AD-*CYH2* destination vectors[9], generating Gal4 DNA binding domain (DB)-viORF hybrid proteins and Gal4 activation domain (AD)-viORF hybrid proteins, respectively.

For expression in mammalian cells, sequence validated viORFs in pDONR223 were recombined into either the Gateway destination vector MSCV-N-Flag-HA-IRES-PURO (NTAP) or MSCV-C-Flag-HA-IRES-PURO (CTAP) (gifts of M. Sowa and J. W. Harper)[31] using LR

Clonase (Invitrogen). The following primer sets were used for resequencing of the resulting plasmids:

MSCV-N and MSCV-C Fwd: 5'-CCCTTGAACCTCCTCGTTCGACC-3'

MSCV-N Rev: 5'-GCCAAAAGACGGCAATATGGTGG-3'

MSCV-C Rev: 5'-GTCGGGCACGTCGTAGGG-3'

Adenovirus: Nine full length ORFs (Supplementary Table 1) were PCR amplified from Ad5 genomic DNA (Human adenovirus 5 strain Adenoid 75 from the ATCC (ATCC VR-5) to generate Gateway entry clones.

Epstein Barr Virus (EBV): Eighty-one EBV ORFs were investigated (Supplementary Table 1). EBV Gateway entry clones and Y2H DB-X and AD-Y clones were previously described[4]. Selected EBV ORFs (Supplementary Table 1) were transferred to the NTAP vector for subsequent transduction of IMR-90 cells.

Human Papillomaviruses (HPV): Seven HPV types[32-38] (virus classifications from http://pave.niaid.nih.gov/#prototypes?type=human) were chosen for this study: HPV6b, 11, 16, 18 and 33 of the alpha genus associated with mucosal lesions (gifts of Y. Jacob), and HPV5 and HPV8 (gifts of Y. Jacob and H. Pfister) of the beta genus which infect cutaneous epithelia. ORFs clones encoding the early region proteins (E4, E5 proteins where relevant, E6, and E7) of all seven HPVs (Supplementary Table 1) were amplified by PCR and cloned by recombination into Gateway vectors.

HPV16 and HPV18 E6 variants: Due to the occurrence of internal splicing in HPV16-E6 and HPV18-E6 and the desire to assure that full-length E6 proteins were expressed in IMR-90 cells, various splice-defective derivatives of HPV16-E6 and HPV18-E6 were also generated. HPV11, 6b, 5 and 8 do not encode internally spliced versions of E6. HPV16-E6X, HPV16-E6XX and HPV18-E6X correspond to the previously described major splice variants of these proteins[39,40]. "E6X" designates the E6*, whereas E6XX designates the E6** splice variants. Primers used to clone and mutagenize HPV16-E6 and HPV18-E6 variants are listed (Supplementary Table 13).

HPV16-E6X, HPV16-E6XX and HPV18-E6X were cloned by PCR amplification using as template genomic DNA from IMR-90 cell populations that had been transduced with NTAP-HPV16-E6 or NTAP-HPV18-E6 expression vectors. These cell lines had been shown to contain integrated copies of these splice variants during the quality control process of the IMR-90 cell culture pipeline. PCR products with the 5' CACC overhang were directionally cloned into the pENTR vector, sequenced, cloned into the NTAP vector and re-sequenced.

To generate versions of HPV16 and HPV18-E6 proteins that do not undergo internal splicing events (designated E6NOX), the splice donor site(s) encoded within the respective E6 ORFs were eliminated by site-directed mutagenesis (QuikChange®; Stratagene)[40]. As a consequence the HPV16-E6NOX and HPV18-E6NOX proteins carry a V to L mutation at amino acid residue 42 and 44, respectively[40].

The previously characterized HPV16-E6 mutants defective in p53 binding (Y54D) and UBE3A/E6AP binding (I128T)[41] were generated by site directed mutagenesis (QuikChange®; Stratagene) of the NTAP-HPV16-E6NOX vector.

PDZ protein binding defective HPV16-E6 and HPV18- E6 mutants were generated by cloning the HPV16-E6NOX and HPV18-E6NOX ORFs into the CTAP vector. This construct fuses HA and FLAG epitope tags to the C-terminal PDZ binding domains and thereby blocks association with cellular PDZ proteins.

Polyomaviruses: ORF clones were obtained from nine polyomaviruses: BK, HPyV6, HPyV7, JCCY, JCMad1, MCPyV, SV40, TSV and WU. The entire early region of BK was PCR amplified from BKPyV Dunlop genomic DNA cloned into pBR322 (gift from Michael Imperiale and Peter Howley). Upon quality control we found that the vector encoded a truncated form of Large T protein. The entire early region of HPyV6 was PCR amplified from pHPyV6-607a (Addgene plasmid 24727)[42]. The entire early region of HPyV7 was PCR amplified from pHPyV7-713a (Addgene plasmid 24728)[42]. The entire early region of JCCY was amplified from JCCY genomic DNA (gift from Igor Koralnik). The entire early region of JCMad1 was amplified from JCMad1

genomic DNA (gift from Igor Koralnik). The entire early region of TSV was amplified from pUC19-TSV[43]. The entire WU early region was amplified from pcDNA WUER (gift from David Wang). SV40 ORFs encoding Large T antigen (LT)[44], Small T antigen (ST)[45], Agnoprotein, 17KT, VP1, VP2, and VP3 were PCR amplified from SV40 genomic DNA.

MCPyV-LST was PCR amplified from Addgene plasmid 24729 pMCPyV-R17a containing the MCPyV R17a genome[42]. MCPyV ORF clones MCPyV-LT, MCPyV-LTtrunc (which contains a premature stop codon at position 961-963), and MCPyV-ST were subcloned from a MCPyV-LST ORF that was first PCR amplified from tumour samples. MCPyV-ST was PCR amplified directly, whereas MCPyV-LT and MCPyV-LTtrunc were obtained by overlap extension PCR[46] to generate a cDNA for LT (nt 429 to 861 of MCV). MCPyV-SPLT was created similarly but with MCPyV-LT as template and with a different set of overlapping primers (spanning nucleotides 1622 to 2778 of MCV). LT, ST, and SPLT use the same Gateway attB1-ORF primer and LT and SPLT use the same Gateway attB2-ORF primer. After two separate PCRs with the relevant Gateway-ORF primer and internal spanning primer pairs, overlap extension PCR produced the final PCR product for subsequent BP cloning to generate MCPyV-LT, MCPyV -LTtrunc, and MCPyV -SPLT entry clones.

B. Yeast two-hybrid (Y2H) assay

Y2H assays used viral ORFs or ORF fragments against the human ORFeome v5.1 collection consisting of ~15,000 full-length human ORFs[10,30] (http://horfdb.dfci.harvard.edu/). Yeast strains Y8800 and Y8930[47], of mating type *MAT***a** and *MAT*α respectively, harboring the genotype *leu2-3,112 trp1-901 his3*Δ*200 ura3-52 gal4*Δ *gal80*Δ *GAL2::ADE2 GAL1::HIS3@LYS2 GAL7::lacZ@MET2 cyh2*^{R}, were transformed together with AD-viORF and DB-viORF constructs, respectively. For Y2H screening *MAT***a** and *MAT*α haploid yeast strains carrying the corresponding AD-viORFs and DB-viORFs were mated against haploid yeast cells of the appropriate mating type carrying DB-human hybrid proteins (DB-huORF) or AD-Human hybrid

proteins (AD-huORFs), respectively. All AD-viORF and AD-huORF plasmids also carry the counter-selectable marker *CYH2,* which allows selection on plates containing cycloheximide (CHX) of yeast cells that do not contain any AD-Y plasmid[9,48]. Colonies growing on CHX-containing medium lacking histidine are considered latent autoactivators and the presumptive pair removed from the dataset[9,48]. At each stage of the interactome mapping pipeline, reporter gene activity is evaluated in parallel both on regular selective plates and on CHX-containing plates.

For Y2H screens with viORFs as DB-hybrid proteins, the 123 Y8930:DB-viORF yeast strains (Y8930 transformed with a unique DB-viORF) were individually mated against mini-libraries containing 188 different Y8800:AD-HuORF yeast strains, each carrying a unique AD-huORF[49]. Twenty-seven of the viORFs consistently behaved as autoactivators when screened as DB-viORF fusions in yeast grown on selective medium (Supplementary Table 14).

Primary reciprocal screens of ~15,000 DB-huORFs against AD-viORFs used individual Y8930:DB-huORF strains mated against virus-specific mini-libraries of Y8800:AD-viORF pools. Primary screens in both orientations were completed twice and all initial positive pairs from the primary screens underwent secondary phenotyping prior to determining the interaction pairs by sequencing[11,49,50]. Each interacting pair was individually retested from fresh stocks of viral and human yeast strains in both orientations, irrespective of the original Y2H orientation in which the viral-human interactions were initially identified. Autoactivating baits were removed at each step of the primary screening and during retesting.

Viral-human interactions found in multiple primary screens and verified by pair-wise retesting are valid biophysical interactions[11,49,50]. Multiple pair-wise retesting can demonstrate that interactions are reproducibly reliable even though any given pair may not retest positive each and every time, providing an additional level of confidence in the data set. Besides the pair-wise retesting done at the time of the original screens, we also carried out an additional retest of all verified Y2H interaction pairs followed by resequencing of all DB-X and AD-Y ORFs

to confirm ORF identity in each interaction pair. The final set of viral-human interactions, consisting of all sequence-verified pairs that successfully retested in a CHX-sensitive manner, contains 53 viral ORFs engaged in 454 interactions with 307 human proteins (Supplementary Table 2). These 454 interactions constitute a set of biophysically real interactions that can be successfully used for further analysis and downstream investigations.

For screens with HPV proteins, all individual viral-human pairs from each HPV strain were also retested in both orientations using the orthologous ORFs of all seven HPVs against all human proteins found by at least one HPV (*i.e.,* a particular huORF found to interact with only HPV16-E6 was tested against all seven HPV E6 proteins multiple times and in both orientations). The final dataset of HPV-human interactions includes those additional interaction pairs even if they were not initially found in the primary screens, as long as they did score positive multiple times in pair-wise retests, in either orientation, without exhibiting growth on CHX-containing medium and after the identity of the specific HPV protein was confirmed by sequencing.

C. Cell culture pipeline

Generating the IMR-90 cell bank: We obtained the human diploid fibroblast cell line IMR-90 (CCL-186) from the American Type Culture Collection (ATCC). IMR-90 cells were cultured in DMEM (Cellgro) supplemented with 15% FBS (Omega Scientific), 1% Pen Strep (GIBCO), 1% Glutamax (GIBCO) and 1% MEM NEAA (GIBCO). A master cell bank of third passage cells consisted of 27 aliquots that were frozen and stored in liquid nitrogen. A single vial of frozen cells was thawed and used to generate each batch of 26 cell lines in the pipeline (Supplementary Fig. 16).

Generating IMR-90 cell lines expressing viORFs: Recombinant retrovirus was produced in Phoenix cells by co-transfecting (Lipofectamine LTX, Invitrogen) plasmids encoding the viORF, GAG/POL and ENV. To produce cell lines expressing each viORF, a vial of IMR-90 cells from

the bank was thawed and expanded (5.0 x 10^5 IMR-90 cells were seeded onto 10 cm plate per transduction) to allow simultaneous transduction of 26 individual viORFs or controls (Supplementary Fig. 16), and then infected twice with recombinant retrovirus for 5 – 8 hr in the presence of 5 μg/ml hexadimethrine bromide (Sigma) with subsequent selection with puromycin (2 μg/ml).

To permit comparisons between different batches of 26 cell lines, each batch included biological replicates of the GFP control and SV40 LT. Five pipeline batches with a total of 87 unique viORFs led to the establishment of 75 IMR-90 cell lines expressing unique viORFs (Fig. 1b and Supplementary Table 1).

Quality control (QC): IMR-90 cells expressing viORFs were uniformly expanded to generate several vials of cells that were frozen and stored. IMR-90 cell lines expressing viORFs were banked at passages 7, 8, and 10. All cell lines had to pass several Quality Control (QC) steps (Supplementary Fig. 16): semi-quantitative western blot analysis for viORF expression using anti-HA antibodies (HA-11 Clone 16B12, Covance), mycoplasma testing (MycoAlert Kit, Lonza), and sequencing of the integrated viORF from genomic DNA post transduction. Sixty-four of the 75 cell lines initially established passed QC (Fig. 1b and Supplementary Table 1). Following QC, cell lines were processed for microarray analysis and tandem affinity purification.

For the sequencing QC step, genomic DNA was extracted from all cell lines (DNeasy Blood and Tissue Kit, Qiagen) and the viORFs were amplified with LA Taq (Takara), using the following vector-specific primer sets:

MSCV-N and MSCV-C Fwd primer: 5'-CGCATGGACACCCAGACCAGGTC-3'

MSCV-N Rev primer: 5'-TCACGACATTCAACAGACCTTGC-3'

MSCV-C Rev primer: 5'-GTCGGGCACGTCGTAGGG-3'

PCR products were extracted from the gel (QIAquick Gel Extraction Kit, Qiagen) and sequenced with the following primers:

MSCV-N and MSCV-C Fwd primer: 5'-CCCTTGAACCTCCTCGTTCGACC-3'

MSCV-N Rev primer: 5'-GCCAAAAGACGGCAATATGGTGG-3'

RNA isolation and microarray analysis: viORFs-expressing IMR-90 cell lines at passage 9 were seeded into five replicates at a density of 1 x 10^6 cells per 10-cm plates and harvested after 48 hr in RNAlater (Ambion). After isolation of total RNA (RNeasy, Qiagen) RNA integrity was determined using a Bioanalyzer (Agilent). Gene expression was assayed using Human Gene 1.0 ST arrays (Affymetrix) according to standard protocols.

Cell proliferation and senescence assays: We measured the growth rate of a subset of IMR-90 lines expressing the viORFs HPV5-E6, HPV8-E6, MCPyV-LTTRUNC, SV40-ST, MCPyV-ST, HPV6b-E6, HPV18-E6X, C-GFP, N-GFP (two biological replicates), EBV-BGLF3, HPV18-E6NOX, HPV16-E6, and SV40-LT (two biological replicates) between passages 9 and 12. Cells were seeded in triplicate onto six 12-well plates (day 0; 2 x 10^4 cells per well) and cell density was measured by crystal violet assay on days 1, 4, 7, 9, 11, and 14. All cell density values were normalized to the value measured at day 1. For senescence assays, IMR-90 cells expressing viORFs between passages 10 and 12 were seeded in triplicate onto 6-well plates and subjected after 48 hr to a SA-b-gal Senescence Colorimetric Assay (Sigma). Stained cells were overlaid with 50% glycerol and photographed at 10X magnification under a Nikon Eclipse E300 microscope with a Spot digital camera (Diagnostic Instruments, Inc.). Three non-overlapping fields were photographed. The number of SA-b-gal positive cells was determined by examination of the digital images[51].

D. HPV E6 oncoproteins and Notch signalling

The U-2 OS cells used in several of the Notch signalling experiments were cultured in DMEM (Invitrogen) supplemented with 10% FBS (Gemini Bio-products), and 1% Pen Strep (Invitrogen).

Generating pNCMV expression vectors encoding HPV E6 proteins: PCR amplification products encoding HPV E6 proteins were amplified using B-Actin HPV E6 expression vectors as

a template[52]. PCR amplification products with a 5' BamHI and a 3' BglII restriction enzyme cleavage sites were directionally cloned into the BamHI site of the pNCMV vector[52,53], resulting in an in-frame fusion of the FLAG and HA epitope tags at the N-terminus of the E6 coding sequence. This approach was used for all HPV E6 proteins except for HPV18 E6 which has an internal BamHI site, and was cloned into the pNCMV vector using PCR amplification products with BglII restriction enzyme cleavage site at both the 5' and 3'. All constructs were sequence validated.

Transfection, cell lysates, immunoprecipitation, immunoblotting, and antibodies: U-2 OS cells were seeded onto 15 cm plates (2 x 10^6) and transfected with 12.5 micrograms of a control pNCMV vector or of a pNCMV vector encoding the indicated HPV E6 proteins using XtremeGENE 9 (Roche). Cells were harvested 48 hours post transfection. Both viORFs-expressing IMR-90 cell lines and transfected U-2 OS cells were lysed in EBC buffer[54] (50 mM Tris HCl, pH 8.5, 150 mM NaCl, 0.5% NP-40, 0.5 mM EDTA, proteinase and phosphatase inhibitors). Extracts were cleared at 14,000 rpm for 15 min. For immunoprecipitations, equal amounts of cell extracts were incubated for 4 hr at 4°C with 30 µl of anti-HA agarose (Sigma). The beads were washed four times with EBC buffer and resuspended in sample loading buffer (Bio-Rad). Proteins were resolved by SDS-PAGE (Criterion™ TGX™ precast gels, Bio-Rad) and subjected to immunoblot analysis with one of the following commercially available antibodies: anti-p300 (A-300-358A, Bethyl Laboratories), anti-MAML1 (4608S, Cell Signaling Technology or A-300-672A, Bethyl Laboratories), anti-E6AP (H-182, sc-25509, Santa-Cruz Biotechnology), anti-vinculin (V9131, Sigma), and anti-HA (HA-11 Clone 16B12, Covance). After incubation with the appropriate secondary antibodies, antigen/antibody complexes were detected by enhanced chemiluminescence (SuperSignal, Pierce).

Reverse transcription-quantitative PCR (RT-qPCR) analyses: Total RNA isolated from IMR-90 cells expressing viORFs and total RNA isolated (RNeasy, Qiagen) from IMR-90 cells expressing *MAML1* shRNA or control was converted to cDNA with a QuantiTect Reverse

Transcription Kit (Qiagen). This cDNA preparation was diluted five-fold, and 1 µl was used per reaction. Real-time PCR quantification, was done in triplicate, using the Brilliant III Ultra Fast SYBR Green QPCR Master Mix (Agilent technologies), and a Stratagene MX3005 instrument. The qPCR primer pairs for the human genes *HES1* (HP208643), *DLL4* (HP213393), and *MAML1* (HP211129) were obtained from OriGene. *GAPDH* (HP205798) was used as the internal reference standard. We used the $2^{(-_{\Delta\Delta}Ct)}$ method[55] to quantify transcript levels.

RNA interference: For lentivirus-mediated RNA interference, the lentiviral construct targeting *MAML1* (SHCLNG-NM_014757, TRCN0000003353, Sigma) [5'-CCGGCATGATACAGTTAAGAGGAATCTCGAGATTCCTCTTAACTGTATCATGTTTTT-3'] or the control empty vector pLKO.1ps (gift of William Hahn) were transfected into HEK293FT cells according to published protocols[56]. Puromycin selection (2 µg/ml) began two days after infecting IMR-90 cells with lentiviruses and was maintained thereafter. IMR-90 cells transfected at passage 9 with *MAML1* shRNA or control pLKO.1ps vectors were seeded (1 x 10^6) into five replicates (5 x 10-cm plates) and harvested after 48 hr in RNAlater (Ambion). After isolation of total RNA (RNeasy, Qiagen) RNA integrity was determined using a Bioanalyzer (Agilent). Gene expression was assayed using Human Gene 1.0 ST arrays. For siRNA-mediated RNA interference ON-TARGETplus SMARTpool siRNA (Thermo Scientific) targeting *MAML1* (L-013417-00; sequences below), or control ON-TARGETplus Non-targeting siRNA (D-001810-10-20; sequences below) were transfected into cells using Lipofectamine RNAiMAX (Invitrogen). Cells were harvested 72 hours post-transfection.

siRNA J-013417-06, MAML1	GUUAGGCUCUCCACAAGUG
siRNA J-013417-07, MAML1	GGCAUAACCCAGAUAGUUG
siRNA J-013417-08, MAML1	GCAGCUGUCCAUAUAAGU
siRNA J-013417-09, MAML1	UCGAAGACCUGCCUUGCAU
siRNA D-001810-01, non-target	UGGUUUACAUGUCGACUAA
siRNA D-001810-02, non-target	UGGUUUACAUGUUGUGUGA
siRNA D-001810-03, non-target	UGGUUUACAUGUUUUCUGA
siRNA D-001810-04, non-target	UGGUUUACAUGUUUUCCUA

Luciferase Reporter Assays: U-2 OS cells were seeded in triplicate onto 6-well plates (2 x 10^5 cells per well). Each well was transfected with 10 nanograms of pGK Renilla, 0.5 micrograms of pGL2 mHes1-Luc, 0.25 microgram of pcDNA3 2HA-ICN1 (gift of Elliot Kieff) and increasing concentrations (25, 50, and 100 nanogram) of the NCMV vector control or NCMV encoding the indicated HPV E6 proteins using X-tremeGENE 9 (Roche). The total amount of DNA per transfection was brought up to 2 microgram per well by adding salmon sperm DNA. After 48 hours cells were lysed with Passive Lysis Buffer (Promega), and subjected to the Dual-Luciferase Reporter Assay (Promega). Luciferase activity was measured using the LMax II^{384} microplate reader (Molecular Devices), and the Renilla luciferase was used as the internal reference standard.

E. Tandem Affinity Purifications (TAP) followed by Mass Spectrometry (MS)

IMR-90 fractionation and protein extract preparation: For each viORF tested the corresponding IMR-90 cells were seeded into twelve to eighteen 15 cm dishes to produce ~2 to 4 x 10^8 cells by harvesting time. Cells were washed *in situ* with ice-cold PBS, harvested using a cell lifter and collected by centrifugation at 500 x g for 5 min at 4°C. Subcellular fractions were obtained by differential digitonin fractionation[57] with the following modifications. To prepare nuclear and cytoplasmic fractions, cell pellets were resuspended in five pellet volumes of ice-cold digitonin extraction buffer (0.015% digitonin, 300 mM sucrose, 100 mM NaCl, 10 mM PIPES, 3 mM $MgCl_2$, pH 6.8) containing protease inhibitors (Roche) and then incubated end-over-end for 10 min at 4°C. Triton X-100 was added to a final concentration of 0.5% and the cell suspension was vortexed for 10 seconds. Triton X-100 extracted cells were then divided equally between three tubes and centrifuged at 1000 x g for 10 min at 4°C. The supernatants (crude cytoplasmic fraction) were transferred to fresh tubes apart from the nuclear pellets. Both fractions were flash frozen in liquid nitrogen and stored at -80°C. To prepare nuclear and

membrane fractions, cell pellets were resuspended in five pellet volumes of ice-cold digitonin extraction buffer and then incubated end-over-end for 10 min at 4°C. The cell suspension was centrifuged at 500 x g for 10 min at 4°C. The resulting pellet was resuspended in five pellet volumes of Triton X-100 extraction buffer (0.05% Triton X-100, 300 mM sucrose, 100 mM NaCl, 10 mM PIPES, 3 mM $MgCl_2$, pH 7.4) containing protease inhibitors and then incubated end-over-end for 30 min at 4°C. The suspension was divided equally into three tubes and centrifuged at 5000 x g for 10 min at 4°C. The supernatant (crude membrane fraction) was transferred to fresh tubes and flash frozen in liquid nitrogen, apart from the nuclear pellets, which were also flash frozen. Before affinity purification the composition of the cytoplasmic and membrane fractions was adjusted to 150 mM NaCl, 50 mM Tris pH 7.5. The composition of the cytoplasmic fraction was additionally adjusted by adding NP40 to a final concentration of 0.05%.

TAP of viral protein complexes: Viral proteins were purified from 6 to 9 plate-equivalents of nuclear and cytoplasmic (or nuclear and membrane) fractions by sequential FLAG and HA immunoprecipitation[12,58]. Tandem affinity purifications were typically done in batches of 2 to 6 viORF-expressing cell lines and included empty vectors or GFP controls cells. For FLAG purification 40 ml of FLAG agarose bead slurry (FLAG-M2 agarose, Sigma) was added to cellular extracts, which were then incubated for 4 hr at 4°C under constant end-over-end mixing. The FLAG-agarose beads were collected by low speed centrifugation and washed three times with 500 ml of cold FLAG IP buffer (150 mM NaCl, 50 mM Tris, 1 mM EDTA, 0.5% NP40, 10% glycerol and protease inhibitors) and once with 500 ml of HA-IP buffer (150 mM NaCl, 50 mM Tris, 1 mM EDTA, 0.05% NP40, 10% glycerol and protease inhibitors). FLAG-purified complexes were recovered by two successive 30 min elutions with 20 ml FLAG elution buffer (400 μg /ml of FLAG peptide in HA IP buffer) at 4°C under constant vortexing. To prevent FLAG agarose bead carry-over, pooled eluates were filtered through a pre-wetted 0.45 mm filter before the subsequent HA purification step. For HA purification 20 ml of HA agarose bead slurry (HA-F7 agarose conjugate, Santa Cruz) was added to cellular fractions, which were then

incubated overnight at 4°C under constant vortexing. HA agarose beads were collected by low speed centrifugation and washed three times with 200 ml of cold HA-IP buffer. HA-purified complexes were recovered by two successive 30 min elutions in 20 ml HA elution buffer (400 μg /ml of 6xHis-HA peptide in HA IP buffer) at 4°C under constant vortexing. To prevent HA agarose bead carry-over, pooled eluates were filtered through a pre-wetted 0.45 mm filter before analysis. A 15% aliquot of each HA eluate was resolved on 4-12% acrylamide gels run in MOPS buffer (NuPAGE, Invitrogen), and viral and associated host proteins were visualised by silver-stain (Supplementary Fig. 17).

Biological replicates: We repeated the tandem affinity purification process for 144 (95%) of the initial 152 sub-cellular fractions used for the first iteration (TAP1). This biological repeat (TAP2) was initiated several months after the completion of TAP1.

Sample digestion of viral-protein multi-component complexes: Purified viral proteins and associated host proteins were denatured and reduced by incubation at 56°C for 30 min in 10 mM DTT and 0.1% RapiGest (Waters). Reduced cysteines were alkylated by adding iodoacetamide to a final concentration of 20 mM and then incubated at room temperature for 20 min in the dark. Protein digestion was carried out overnight at 37°C after adding 2 mg of trypsin and adjusting the pH to 8.0 with a 1 M Tris solution. For tryptic peptide clean-up RapiGest was inactivated following the protocol of the manufacturer. Tryptic peptides were purified by batch-mode reverse-phase C18 chromatography (Poros 10R2, Applied Biosystems) using 40 ml of a 50% bead slurry in RP buffer A (0.1% trifluoroacetic acid), washed with 100 ml of the same buffer and eluted with 50 ml of RP buffer B (40% acetonitrile in 0.1% trifluoroacetic acid). After vacuum concentration, peptides were further purified by strong cation exchange SCX chromatography (Poros 20HS, Applied Biosystems) using 20 ml of a 50% bead slurry in SCX buffer A (25% ACN in 0.1% formic acid), washed with 20 ml of the same buffer and eluted with 20 ml of SCX buffer B (25% ACN, 300 mM KCl in 0.1% formic acid).

Primary LC-MS/MS data acquisition: After vacuum concentration, half of each sample was loaded onto a pre-column (100 mm I.D. x 4 cm; packed with POROS 10R2, Applied Biosystems) at a flow rate of 4 ml/min for 15 min, using a NanoAcquity Sample Manager (20 ml sample loop) and UPLC pump[59]. After loading, the peptides were gradient eluted (1-30% B in 45 min; buffer A: 0.2 M acetic acid, buffer B: 0.2 M acetic acid in acetonitrile) at a flow rate of ~50 nl/min to an analytical column (30 mm I.D. x 12 cm; packed with Monitor 5 mm C18 from Column Engineering, Ontario, CA), and introduced into an LTQ-Orbitrap XL mass spectrometer (ThermoFisher Scientific) by electrospray ionization (spray voltage = 2200V). Three rapid LC gradients were included at the end of every analysis to minimize peptide carry-over between successive analyses and to re-equilibrate the columns. The mass spectrometer was programmed to operate in data dependent mode, such that the top eight most abundant precursors in each MS scan (detected in the Orbitrap mass analyzer, resolution = 60,000) were subjected to MS/MS resolution (CAD, linear ion trap detection, collision energy = 35%, precursor isolation width = 2.8 Da, threshold = 20,000). Dynamic exclusion was enabled with a repeat count of 1 and a repeat duration of 30 sec. The second half of each sample was reanalyzed by LC-MS/MS after the completion of the initial set of analyses ("Technical Replicates").

Database searching: Orbitrap raw data files were processed within the multiplierz software environment[60]. MS spectra were recalibrated using the background ion $(Si(CH3)_2O)_6$ at m/z 445.12 +/- 0.03 and converted into a Mascot generic format (.mgf). MS spectra were searched using Mascot version 2.3 against four appended databases of: i) human protein sequences (downloaded from RefSeq on 07/11/2011); ii) viral proteins analyzed in this study; iii) common lab contaminants and iv) a decoy database generated by reversing the sequences from the human database. Search parameters specified a precursor ion mass tolerance of 1 Da, a product ion mass tolerance of 0.6 Da, trypsin with a maximum of two missed cleavages,

variable oxidation of methionine (M, +16 Da), and carbamidomethylation of cysteines (C, +57 Da).

Mascot search results from all experimental runs were separately imported for TAP1 and TAP2 into two multiplierz mzResults files (.mzd)[61] for further processing: The list of peptide hits from the Mascot searches was filtered to exclude peptides with precursor mass error greater than 5 ppm. Sequence matches to the decoy database were used to establish a 1% false discovery rate (FDR) filter for the resulting peptide identifications. A rapid peptide matching algorithm[62] was then used to map peptide sequences to all possible human Entrez Gene IDs. Only peptides that were uniquely assignable to a single Entrez Gene ID were considered as detection evidence. For each viral-host protein association, peptide evidences for a given target human protein were combined across different subcellular fractions and LC-MS/MS technical replicates, when applicable. Any Entrez Gene ID present in the "Tandome" (Supplementary Table 15) was automatically removed from the list of putative interactors.

TAP-LC-MS quality controls: We used a multi-tiered approach to minimize the occurrence of peptide carryover across LC-MS analyses: 1) samples within each batch of LC-MS analyses were injected in order of increasing abundance and complexity, as evaluated by SDS-PAGE/silver stain visualisation of the corresponding TAPs; 2) three rapid LC gradients were run at the end of every LC-MS analysis; 3) for all LC-MS analyses, we systematically evaluated run-to-run carry-over by calculating extracted ion chromatogram (XIC) intensities of peptides derived from viral proteins. 4) samples within each purification batch, and LC-MS analysis sequence were randomized between biological replicates (TAP1 and TAP2). This multi-step quality control process led us to discard all LC-MS data for EBV-BFRF2 and EBV-BRRF2 from our TAP2 (and as a consequence from the final dataset, since we only report proteins identified reproducibly across TAP1 and TAP2), and for two GFP negative controls. Because we identified cross contamination between HPV8 E6 and HPV11 E6 at an early point of our project, we were able to repeat the TAP and LC-MS analyses to generate a "clean" dataset for bio-replicates.

Pathway visualisation and analysis: Protein associations detected between HPV-E6 and HPV-E7 oncoproteins and host proteins were imported into Pathway Palette for data analysis and visualisation[63]. For each of these datasets, we first calculated the probability that the observed set of human proteins targeted by multiple E6 or E7 across different HPV types (HPV5, HPV6b, HPV8, HPV11, HPV16 and HPV18) could occur by chance (Fig. 1d and Supplementary Fig. 4): for each viral protein of a given HPV type, the same number of associated human proteins as that experimentally observed in our data was randomly chosen from our HI-2 database. This process was repeated 1,000 times. The number of human proteins associated with two or more viral proteins from different HPV types was calculated in each simulated network and the frequency distribution of random graphs with a given number of shared nodes was computed. Lastly, this distribution was used to calculate the probability that the number of human proteins targeted by multiple HPV types observed in our experimental graphs (40 for E6 and 25 for E7) could occur by chance. We also calculated the probability of detecting human proteins targeted by two viral proteins from HPVs of the same class: the total number of human proteins targeted by viral proteins from pairs of HPVs of the same class ([HPV5 and HPV8], [HPV6b and HPV11], [HPV16 and HPV18]) was extracted from our TAP-MS data. This value was divided by the total number of human proteins that could theoretically be shared across all pairs of HPV from the same class in our observed networks. The same ratio was calculated for human proteins targeted by two viral proteins from HPVs of different classes. The ratio of these two ratios (4.9 for E6 and 1.05 for E7, indicated by green arrows in the frequency distribution on the right) represents the overall enrichment of human proteins targeted by proteins encoded by HPVs of the same versus different class observed in our graphs. One thousand randomized networks were created as before and used to evaluate the probability that the observed ratio could occur by chance (Fig. 1d and Supplementary Fig. 4). Graphs of protein interaction networks for each viral protein, including the silver stain image of a representative

tandem affinity purification, can be accessed at: http://blaispathways.dfci.harvard.edu/Palette.html

F. Subtracting likely non-specific protein associations ("Tandome")

Twenty five control TAP experiments were run at multiple points during the entire project. These controls consisted of FLAG-HA purifications from nuclear, cytoplasmic or membrane extracts prepared from parental IMR-90 cells, from IMR-90 expressing FLAG-HA-tagged GFP or from IMR-90 cells transduced with the empty NTAP or CTAP vectors. An additional 83 control TAP experiments consisted of FLAG-HA or FLAG-StrepTactin purifications on HeLa subcellular extracts. Any Entrez Gene ID identified in more than 1% of the control TAP experiments (whether or not the Gene ID was detected by a uniquely assignable peptide sequence) was included in the "Tandome" list. Any of the 679 Gene IDs present in the Tandome was systematically removed from all viral protein interaction TAP datasets (Supplementary Table 15).

G. TAP reproducibility

The reproducibility rate of TAP-MS across the two biological replicates (TAP1 and TAP2) corresponds to the percentage of viral-host protein associations detected in one TAP (the "reference TAP") at a given uniquely assignable peptide threshold, which was confirmed in the other (the "repeat TAP"). This percentage was calculated for both permutations in which the "reference TAP" and the "repeat TAP" were successively (TAP1 and TAP2) and (TAP2 and TAP1), respectively. The average reproducibility rate was also calculated. As expected, reproducibility increased as a function of the number of unique peptides detected (Supplementary Fig. 13).

H. Virus-human Positive Reference Set (PRS)

The VirusMINT database curates from the literature viral-host interactions for 10 disease-relevant virus groups[64]. Each curated interaction gets an evidence count, where multiple evidences could be multiple publications, or multiple methodological verifications, reporting the same interaction. To generate a Positive Reference Set (PRS) of high confidence interactions, we only collected interactions supported by two or more evidences in VirusMINT, resulting in a master list of 134 interactions. Given the different types of associations detected by Y2H and TAP-MS, a specific PRS was designed for each one of these datasets (Supplementary Table 16). For the Y2H-specific PRS, we first selected host proteins with an available ORF clone in the human ORFeome 5.1 collection (ORFeome 5.1), resulting in a list of 94 viral-host protein interactions. This list was filtered to remove interactions involving known autoactivators in the yeast two-hybrid assay (Supplementary Table 14) and interactions that were not tested in the yeast two-hybrid screen. The resulting Y2H-specific PRS contains 62 interactions. Starting with the master list of 135 interactions, the TAP-specific PRS was created by removing interactions that involved viruses that were not tested in the TAP-MS assay and interactions that involved a human protein present in our "Tandome" (Supplementary Table 15), resulting in a TAP-specific PRS of 94 interactions (Supplementary Table 16).

I. Pathway enrichment analysis

Pathway enrichment used FuncAssociate 2.0 (http://llama.mshri.on.ca/funcassociate/)[65], which performs a Fisher's Exact Test and compares the observed *P*-value for the query set to the minimum *P*-value obtained from Monte Carlo sampling of the appropriate gene sample space. When the query set of genes arose from multiple sample spaces, a Python script of the FuncAssociate approach was used to independently sample from each space so that the random query sets matched the original set in size and composition. KEGG pathway annotations were downloaded from http://www.genome.jp/kegg/download/. Gene Ontology

annotations were downloaded from the FuncAssociate 2.0 website. For all analyses we considered only specific pathways and functional terms, defined here as those which have been annotated with fewer than 1000 genes. All *P*-values were adjusted for testing of all pathways in the GOA or KEGG databases (Supplementary Table 17), including the reported *P*-values for apoptosis pathway members. Additionally, pathway enrichment of TAP-MS data sets were assessed at unique peptide thresholds ranging from 1 to 5, with the resulting *P*-values corrected for the 5 hypotheses tested, using the Bonferroni method. Odds ratios were calculated using 2x2 contingency tables.

J. Microarray preprocessing and differential expression

Microarray data analysis used R/Bioconductor (http://www.bioconductor.org). Data was first normalized by robust multiarray averaging (RMA) using a custom Chip Description File (CDF) from the Michigan Microarray Lab (http://brainarray.mbni.med.umich.edu, Version 13). The Human Gene 1.0 ST array from Affymetrix contains 764,885 probes, with 26 probes located across the full length of each annotated gene from the March 2006 human genome sequence assembly. The Michigan Microarray Lab CDF reanalyzes and matches each probe sequence to the most current gene annotation, producing summarized expression values for 19628 genes with Entrez Gene IDs. Using this custom CDF rather than the original Affymetrix annotation ensured a more accurate estimate of gene expression levels[66].

A total of 435 microarrays were used to profile cell lines expressing 63 viORFs, as well as controls. The arrays were processed in eight groups (Supplementary Table 18). Note that each group includes cell lines expressing proteins from a variety of viruses, five of the groups include control cell lines, and the order of the arrays was randomized. These preemptive measures were taken in order to eliminate potential systematic biases that could affect our results and interpretation. The normalized data showed a slight propensity to cluster according to the batch covariate (Supplementary Fig. 18). We used ComBat[67] to reduce this batch effect. ComBat

applies a linear model to the expression levels, with factors corresponding to global expression, sample-dependent expression, additive/multiplicative batch effects and noise. It then synthesizes information across genes using an empirical Bayes framework, and thus estimates and removes batch effects (Supplementary Fig. 18). The R package *limma* was used to test for differential expression between each of the 63 conditions and GFP-control arrays. Multiple testing adjustments were carried out with the Benjamini-Hochberg method.

K. Microarray gene and sample clustering

We used the R package *limma* to test for differential expression across all pairwise comparisons between biologically distinct samples, including the control. The top 2,944 genes were selected by ranking all genes on the array by the number of significant comparisons ($P_{adj} < 0.005$; Supplementary Table 19). We applied the R package *mclust* to carry out model-based clustering on the top 2,944 genes. We used the *hmap* function in the *seriation* R package to generate the heatmap representation of mean fold-change of expression of genes within each cluster.

L. Predicting cell-specific transcription factor binding sites

Position weight matrices (PWMs) corresponding to mammalian transcription factors were obtained from the TRANSFAC 7.0, Jaspar, and UniProbe databases. We supplemented available PWMs by downloading experimentally determined genome-wide transcription factor binding sites from the GEO and the UCSC genome databases. We conducted *de novo* motif discovery using the MEME 4.3 and Weeder 1.4.2 software tools. We then mapped motifs to genes using either the keys provided by the databases or based on the antibody used for the ChIP experiment. A total of 610 genes were identified for which at least one PWM was available. Of these 361 had at least one “high-quality” PWM, defined as being derived from either a ChIP-seq or ChIP-chip experiment or from a universal protein binding microarray (PBM).

Database	Number of PWMs
Transfac: http://www.biobase-international.com/	709
Jaspar: http://jaspar.genereg.net/	721
Uniprobe (not in Jaspar): http://the_brain.bwh.harvard.edu/uniprobe/	26
UCSC: http://genome.ucsc.edu/ or GEO: http://www.ncbi.nlm.nih.gov/geo/	67

Clustering of position weight matrices: Our identification of regulatory variants depends on correctly mapping genes to motifs, although our integrative approach can tolerate some errors. We took a computational approach to improving the quality of the gene to motif mapping. For genes for which a "high-quality PWM" was available, we eliminated any other PWM that showed no resemblance. To do so systematically, we hierarchically clustered all PWMs. First 100,000 random 30-mers were generated. Then for each of the ~1500 PWMs, the maximum match score (MMS) for overlap between the motif and 30-mer was generated by comparing the observed frequency f_{bj} of base b at each position j in the PWM with the background frequency of that base p_b in the genome and summing over the length of the motif:

$$MMS = -\sum \log_2 (f_{bj} / p_b)$$

The matrix of 30-mer x PWM was hierarchically clustered using the *hclust* function in R, with average linkage and the Pearson correlation coefficient used as the metric for similarity. A coarse clustering cutpoint (height = 0.85, resulting in 232 total PWM clusters) was used to broadly group together PWMs with some evidence of similarity. For genes having at least one high quality PWM, we eliminated all PWMs that failed to cluster with any of these (although we retained PWMs that represented dimers of the high quality motif). This approach rejected 113 gene-PWM pairs.

Selection of motifs for genome-wide binding site prediction and enrichment analysis: We selected individual motifs for subsequent analysis with the goal of providing at least one high

quality motif for every gene. For each of the 610 genes we selected all motifs derived from a ChIP-seq, ChIP-chip or protein binding microarray experiment. If any other motif existed (*i.e.* from Transfac or Jaspar), we included it if it appeared to correspond to a dimer or half-site not represented in the otherwise selected high-quality motifs. All motifs were selected prior to further analysis for genes without any high-quality PWM. A total of 841 motifs corresponding to 573 TFs satisfied the above criteria and the minimum match score of 12.

M. Transcription factor binding site (TFBS) enrichment analysis

IMR-90-specific TFBS were predicted by identifying binding sites with an MMS≥12 within accessible chromatin regions. Accessible chromatin regions were identified using IMR-90-specific genome-wide DNase-seq data downloaded in wig file format from the GEO database (GSM530665, GSM530666). Wig files were processed using the BEADS packages[68] to correct for systematic biases in GC-content and ability to map reads. Corrected counts were smoothed and binned into 250 bp groups. Bins with a significant number of counts were identified by assuming that such counts follow a Poisson distribution. The Benjamini-Hochberg procedure was used to estimate false discovery rates and contiguous bins with FDR < 0.001 were merged into peaks. Only peaks observed within both replicate IMR-90 data sets were used for subsequent analyses.

The enrichment of TFBS within a given expression cluster was computed for each TF by counting the number of TFBS in DNase accessible regions located within 25 kb of the transcription start site of any transcript within this cluster and comparing the count to a null distribution of TFBS from length- and GC-content matched sets of mappable genomic regions. For each cluster, nearby DNase-seq peaks were assembled into a single file and the peaks trimmed to common lengths of 250, 500, 750, 1000, 1250, 1500, 1750, 2000, 2250 or 2500 bp. Next, 200 sets of length-matched “control” DNase-accessible segments were sampled from the random genome segments. For each motif, the number of TFBS for each set of 200 sequences

was computed, and the values fitted to a Poisson distribution. This null distribution was used to estimate the *P*-value for the actual number of TFBS within IMR-90 cells. The enrichment *P*-values for all TF motifs within each cluster were combined and a tail-based false discovery rate was estimated using the Benjamini-Hochberg procedure[69]. TFs with motifs significant at an FDR of 0.5% or less were selected for further analysis. In total, we found 239 TFs to be enriched for target genes in one or more clusters.

N. Randomization of gene clusters for enrichment analyses

To assess the biological coherence of identified gene clusters (Supplementary Tables 5 and 6) we computed the random expectation of enriched KEGG pathways and GO terms. For 100 iterations, we randomly reassigned the 2,944 genes to 31 clusters matched in size to those observed and repeated the enrichment analysis, counting at each iteration the average number of significant KEGG pathways and GO terms per cluster. The computationally more intensive GO term enrichment was performed by counting the average number of nominally significant pathways ($P < 0.05$) per cluster. We found that there were more GO and KEGG annotations than expected by chance ($P < 0.01$), underscoring the functional coherence of these clusters (Supplementary Fig. 19).

O. Evaluating predictive power of regulatory cascades

We computed scores for the viORF-TF-cluster network in the following way: for each viORF, we first identified all genes for which we would expect differential expression. To do so, we identified all TFs that had a physical association with the protein encoded by the viORF or was differentially expressed in response to its expression, considering fold changes greater than 1.2 at a $P_{adj} < 0.001$. For each of these TFs, we then identified all microarray clusters enriched for genes containing the corresponding binding site within their promoters. Within these clusters, all genes with a high-probability binding site for that TF in their promoter were selected. The union of these genes across all TFs targeted by a given viORF represents the set of genes that would

be expected to show differential expression upon transduction of this viORF. We chose as the meaningful statistic the fraction of these target genes that were observed to be differentially expressed on the corresponding array (*i.e.* where that viORF was transduced) and then computed the average over all viORFs. A null distribution was created by computing the same score for 1,000 random permutations of the 63 array conditions (Fig. 2b). We used Cytoscape 2.8.1 to visualise the network[70]. The resulting network (Supplementary Fig. 9) illustrates how viral proteins may regulate DNA damage response, as well as core functions of fibroblasts, like proliferation and cell adhesion via multiple TFs.

P. Building a probabilistic map of IMR-90 specific RBPJ binding sites

To build a map of IMR-90 specific TF binding sites, we followed a previously described probabilistic framework[71]. The approach combines PWM information with complementary genomic and experimental data in a mixture model, and outputs a posterior probability of DNA binding. Features of interest include chromatin accessibility (as obtained through genome-wide DNase-seq number or histone mark ChIP-seq experiments), multi-species sequence conservation, and distance from the nearest gene transcription start site.

We downloaded publicly available data from diverse sites. FASTA files corresponding to the primary sequence of the complete human genome (hg19) were downloaded from the UCSC FTP site. IMR-90-specific chromatin accessibility data was generated as in “Transcription factor binding site (TFBS) enrichment analysis”. We collected individual base conservation information from the phastCon tables in the UCSC genome browser. To identify transcription start sites for all RefSeq IDs we downloaded the refGene.txt file from the UCSC Genome Browser FTP site. Using the data obtained for each of the RBPJ PWM, we proceeded as follows:

1. Match score: the complete genome was scanned to compute a match score at each base pair. All matches with a score ≥12 (an arbitrary threshold for similarity) were included as potential binding sites.

2. Chromatin accessibility score: we defined a 200 bp “chromatin accessibility window” around each potential binding site from (1) by assigning the total number of DNase-seq reads in the 100 bp upstream and downstream of the motif starting position.
3. Nearest transcription start site: we identified the nearest transcription start site to each of the potential binding sites.
4. Conservation score: we averaged the phastCon conservation score over the length of the motif and assigned this average score to each potential binding site.

We then used the CENTIPEDE R package[71] to infer an IMR-90-specific probability of RBPJ binding at each position in the genome.

To generate a short list of Notch pathway transcriptional targets, we took the union of Notch pathway members (as defined in KEGG) and potential Notch target genes[72]. From this set, we chose the genes containing an IMR-90-specific high probability RBPJ binding site in their promoters to generate the heatmap in Fig. 3b.

Q. Notch pathway enrichment analysis

Enrichment of viral protein targets for Notch pathway members was performed by comparing the number of co-complex associations or direct interactions between viral proteins and Notch pathway members (as defined by KEGG Pathways) to the number of interactions selected at random for each viral protein. The sampling methods have been defined in “HI-2 and analysis of overlaps between Y2H, TAP-MS and PRS”.

R. Identification of loci implicated in familial and somatic cancer

The COSMIC Classic list of causal genes with somatic mutations observed in tumours was downloaded from http://www.sanger.ac.uk/genetics/CGP/Classic/. To assemble the genome-wide association loci associated with cancer we consulted the publicly available Catalog of Published Genome-Wide Association Studies (http://www.genome.gov/gwastudies/) and selected all variants associated with cancer. We grouped variants showing evidence of linkage

disequilibrium and, for the resulting 107 loci, selected as the index SNP the most highly associated variant for any phenotype.

We next mapped SNPs to neighboring genes, a task for which there is no consensus procedure[73]. We first defined the boundaries in which SNPs tagged by the index variant may reside. To do so, we identified all SNPs that were in close linkage disequilibrium ($R^2 \geq 0.5$) with the index SNP. We then marched outward from the outermost SNP positions to the nearest recombination hotspot[74] (downloaded from the UCSC genome browser). We assumed that any SNP within the region bounded by the two recombination hot spots may be a causal variant responsible for the original association. Since variants within enhancers can act at considerable distance from their physical position, we also considered all genes within 250 kb from the hotspot boundaries. When no genes resided within these boundaries, we moved outwards in 50 Kb intervals until a gene was identified (Supplementary Table 20).

To compile SCNA loci implicated in somatic forms of cancer we consulted a recent compendium[27] of loci derived from SNP arrays of over 3,000 tumour samples. Physical boundaries for the SCNAs were used to map each locus to genes contained fully within the SCNA interval (Supplementary Table 21).

S. Testing enrichment of gene sets for cancer genes

We followed the simple hypothesis that the viral interactome is enriched in cancer genes and looked for overlap between viral interaction partners and cancer genes highlighted in the COSMIC Classic list (Supplementary Table 8). The number of overlapping genes was compared to that obtained from 10,000 random gene sets matched in size and composition to the query set.

T. Viral target overlap with candidate cancer genes identified by transposon screens

We identified four “Sleeping Beauty” transposon-based murine insertional mutagenesis screens focused on finding genetic determinants of tumourigenesis in murine cancer models. These

studies correspond to murine models of colon cancer with[75,76], or without[77] mutations of the *Apc* tumour suppressor, and pancreatic carcinoma[78]. The list of candidate cancer genes was extracted from the main text or supplementary data of each manuscript and pooled to form a single set of 1,359 genes (Supplementary Table 10). As expected, the candidate genes identified through insertional mutagenesis exhibited a marked bias towards long genes. We also noted that genes in this candidate set tend to be highly expressed based on our IMR90 microarray data.

We assessed the overlap between the Sleeping Beauty candidate genes and the VirHost set by permutation. The analysis was limited to the universe of human genes with mouse orthologues (downloaded from the Mouse Genome Database). We corrected for the bias observed in gene expression and gene length by binning genes into quartiles based on mRNA expression and based on gene length using the longest distance between the transcription start site and 3' end of all transcripts corresponding to a single gene. A total of sixteen bins were created, corresponding to all combinations between bins of expression and gene length. 1,249 Sleeping Beauty candidate genes and 900 VirHost genes could be assigned to one of these bins, with 156 genes common to the two sets. To assess the chance expectation for overlap, random sets of 1,249 genes were selected that matched the original query set in gene length and expression composition. Ten thousand iterations were used and the empirical *P*-value determined to be the fraction of random gene sets with at least 156 overlapping genes.

We also identified an insertional mutagenesis screen performed to find genetic determinants of leukemia using the Maloney Murine Leukemia virus[79]. The candidate gene set was downloaded from the Supplementary Data, mapped to human orthologues and tested for overlap against the VirHost set. A significant overlap (20 out of 213 genes, $P = 0.048$) was found.

U. Comparison of viral interactome and prioritisation of cancer genes

Catalogues of somatic mutations from genome-wide cancer sequencing studies were obtained from the supplementary tables of nine recently published studies[80-88]. The deleteriousness of each mutation was assessed using the Polyphen2 web server (Polyphen2), which provides a score (HumVar) between 0 and 1, with 1 being the most damaging (Supplementary Table 11). A cumulative deleteriousness score for each protein was obtained by summing all Polyphen2 scores. We observed that large proteins tended to have higher scores and thus normalized scores for protein length. All proteins with at least one somatic mutation were ranked by cumulative deleteriousness score, with TP53 having the highest score (Supplementary Fig. 14). For a direct comparison of COSMIC Classic gene overlap with the viral interactome at various peptide count thresholds, a matching number of somatically mutated proteins were selected from the top of the ranked list and statistical significance assessed by Fisher's Exact Test. Genes at SCNA-AMP, SCNA-DEL and GWAS loci were also tested against COSMIC Classic genes using Fisher's Exact Test. Odds ratios were calculated with 2x2 contingency tables.

Supplementary Notes

1. Significantly targeted and untargeted proteins

"Significantly targeted or untargeted" host proteins are sets of proteins that interact with viral proteins at a higher or lower frequency, respectively, than would be expected given their degree in the human HI-2 interaction network. The full set of viral-human Y2H interactions (Supplementary Table 2) includes interactions obtained after recursive testing of the viral proteins from all seven HPV types against a human interactor found by any one HPV in the initial screen. To eliminate any bias introduced by the recursive nature of the verification retest experiments with HPV proteins, and insure that all viruses are analyzed within the context of the same reference search space this analysis only included interactions detected in the primary screens and directly verified by pairwise retesting. In addition, we further restricted this analysis to human proteins present in HI-2 (174 out of 307). We first counted the number of interactions between viral proteins and human targets in the virus-human interaction network and then randomly selected an identical number of proteins from HI-2. Sampling was performed with replacement (*i.e*. a protein could be chosen more than once) and the likelihood of choosing any protein determined by its number of interacting partners (degree) in HI-2. This process was repeated 10,000 times to generate a random distribution When less than 5% of the simulations generated a number of interactions equal to or greater than the number of experimentally observed interactions for a given host protein, this was considered "significantly targeted". When more than 95% of the simulations generated a number of interactions equal to or greater than the number of experimentally determined interactions, the host protein was considered "significantly untargeted". Otherwise, the human protein was considered "expectedly targeted".

2. HI-2 and analysis of overlaps between Y2H, TAP-MS and their respective PRS

The human protein-protein interactions previously reported in our three high-throughput datasets[11,49,50] were updated to the gene annotations in human ORFeome 5.1 and then

combined to generate the HI-2 list of unique interactions. Distinctions between ORFs corresponding to multiple splice variant of a human gene were ignored when reporting protein-protein interactions.

The significance of overlaps between different datasets was assessed by Fisher's Exact Test or permutation. We examined three overlaps, the overlap between TAP-MS and Y2H, as well as the overlap of these two datasets separately with their respective PRS (Supplementary Table 16). The search space for calculating the Y2H overlap with the Y2H-specific PRS is the hORFeome v5.1. The search space for the TAP-MS and Y2H overlap corresponds to the intersection of the hORFeome 5.1 and TAP-MS search spaces. The TAP-MS search space was used to calculate the overlap between the TAP-MS dataset and the TAP-MS-specific PRS. We noticed that host proteins identified as viral targets by the TAP-MS technology tend to be encoded by genes with high mRNA expression in IMR-90 cells (Supplementary Fig. 20). To circumvent this expression bias when considering the TAP-MS dataset, random sampling was performed from a set of genes matching the query set in terms of expression quartiles, as judged from genome-wide microarray analysis of IMR-90 cells.

Statistical significance of overlaps observed between the Y2H dataset or the TAP-MS dataset and their respective PRS were computed by random sampling from the appropriate search space, and controlling for expression bias in the case of TAP-MS. Both our TAP-MS co-complex and Y2H binary datasets significantly overlapped with PRS pairs (17 out of 94 protein pairs [$P < 0.001$] and 1 out of 62 protein pairs [$P = 0.047$] respectively). The overlap between our co-complex viral-host protein associations and those in the PRS varied as a function of analytical parameters in our TAP-MS data: We observed that the fold enrichment of PRS proteins in the set of host proteins detected in co-complex associations increased with the number of unique peptides, although with fewer associations recovered at more stringent thresholds (Supplementary Fig. 21).

Twenty four viORFs were found to have interactors in both TAP-MS and Y2H datasets (Supplementary Fig. 1), and of these a total of six viORF-host protein interactions were common to the two technologies (Supplementary Fig. 2). The observed overlap of six interactions was significantly higher ($P < 0.001$) than the most common random expectation of 0.81 interactions (Supplementary Fig. 2). To test whether the set of genes identified by TAP-MS as targets of a given viral protein included the immediate neighbours of viral targets identified by Y2H, we repeated the analysis after expanding the list of Y2H targets to their direct neighbours in HI-2. We observed that targets of viral protein identified through both technologies showed a significant tendency to interact with each other ($P < 0.001$; Supplementary Fig. 2) implying that host targets of the two technologies tend to 'fall in the same neighbourhood' of the network.

3. Measurement of the precision of the Y2H dataset by wNAPPA

We measured the fraction of true biophysical interactions in our virus-host binary Y2H dataset (precision) by comparing the recovery rate of the detected interactions to those of the Positive Reference Set (PRS) and Random Reference Set (RRS) in a wNAPPA orthogonal assay. In the wNAPPA assay, bait and prey fusion proteins were expressed by coupled transcription-translation and GST-tagged bait proteins captured by anti-GST antibody. Interactions were detected using antibody to HA with standard immunochemical protocols. One viORF (Adeno5 E1A) was found to bind non-specifically to the GST plate and was therefore removed from subsequent analysis of the wNAPPA data (Supplementary Table 22). The precision was calculated for a recovery rate of 1% for the RRS, corresponding to a z-score threshold of 1.5. At this threshold, we observe recovery rates of 9% ± 1% for the interactions detected by Y2H, 10% ± 4% for the PRS interactions and precisely 1% ± 1% for the RRS interactions. The precision value was calculated according to the following formula[50].

$$I_{+} = \frac{I_{obs} - f_{+}}{1 - f_{-} - f_{+}}$$

where I_+ is the number of true positives in the detected interactions (precision), f_+ is the false positive rate of wNAPPA (obtained as the fraction of RRS pairs reported positive) and $(1 – f_-)$ is the fraction of Y2H supported PRS pairs that report positive in wNAPPA. Error bars of the precision were calculated according to the variation of those values in the [1.5, 1.6] z-score window. These measurements led to a precision of 90% +/- 7% for the virus-host binary Y2H dataset. We determined the fraction of true biophysical interactions in our virus-host binary Y2H dataset to be ~90% by comparing the validation rates of the dataset to those of the PRS and random reference set (Supplementary Fig. 22).

4. Network motif identification

We hypothesized that the striking variation in expression of multiple clusters may relate to TF activity reinforcing feedback or feedforward loops. We thus used the list of predicted TF binding sites to explore the interconnectedness of TFs. We focused on the TFs that were enriched for TFBS in expression clusters and physically associated with or were differentially expressed in response to expression of viral proteins (Supplementary Table 7). We looked for instances of autoregulatory loops, feedback loops, and feed-forward loops – network motifs that might amplify or modulate signals in biological networks – and compared the number observed in our data to that observed after 1,000 random selections of TFs chosen from the set of TFs with one or more predicted high-probability promoter binding site. There were far more interconnections than expected by chance ($P < 0.001$; Supplementary Fig. 6), suggesting that viral proteins target a dense regulatory TF network.

5. Cellular growth phenotypes

If gene expression clusters regulated by viral proteins were relevant to human cancer, then the downstream variation in expression of host genes would be reflected in cellular growth phenotypes. We measured growth and senescence rates for fifteen IMR-90 viral protein expressing and control cell lines, and computed their Pearson correlation coefficients (R) with

the average gene expression level in each cluster. The significance of the correlation was assessed for each cluster by randomly sampling 1,000 sets of equal size from the full gene universe represented on the microarray. The resulting *P* values were adjusted for multiple testing across the 31 clusters using the Benjamini-Hochberg procedure. There was significant correlation between five clusters and growth or senescence phenotypes. Cluster C31 demonstrated the strongest correlation between mRNA expression and both cellular growth (R = 0.8, P_{adj} = 0.008) and senescence (R = -0.92, P_{adj} = 0.016) (Supplementary Fig. 8). This correlation suggests that the expression variation observed in each cluster across cell lines (Fig. 2a) is functionally relevant to viral proteins altering cellular phenotypes.

6. Cancer pathways perturbed by viORFs

Viral proteins perturb diverse core signalling pathways and biological functions of the host cell, including many of the known hallmarks of cancer[14]. Dysregulation of these viORF-perturbed pathways could lead to tumourigenesis. Together, our transcriptional analysis and the interactome datasets shed light on known and potentially novel connections between viral proteins and cancer mechanisms.

We found several signatures of growth suppressor evasion in the transcriptional data. Viral proteins in Group III, which include polyomavirus and high-risk HPV proteins that target RB1, were correlated with increased expression of genes involved in cell proliferation, including components of the cell cycle (CDC25A, CDC25C, CDK1) and DNA replication machinery (PCNA, RFC and MCM family members) and genes involved in nucleosome assembly (HIST1H histone family). TP53-dependent pathways were typically downregulated in Group III proteins, reflecting that many of these viral proteins bind to and inactivate TP53. Pathways include cell-cycle arrest mediated by CDKN1A as well as cell death through BAX and TRAIL pathway components (cluster C12, 2.0-fold enriched for TP53 binding sites, *P* = 2.3 X 10^{-9}). We also observed that Group II viral proteins, which include alternatively spliced E6 proteins from high-

risk HPV types, HPV E5 proteins, E7 proteins from low-risk HPV6b and HPV11 and high-risk HPV16, and many EBV proteins including EBNALP, induced a large decrease in TGFβ signalling genes (cluster C16).

Metabolic dysregulation is often observed in cancer cells. Group III viral proteins induced high expression of genes involved in cholesterol biosynthesis genes (DHCR24, HMGCS1, SQLE) and fatty acid metabolism (MGLL, FABP3, ELOVL6, SCD), potentially reflecting the need of rapidly proliferating cells and tumours for cholesterol[89] and unsaturated fatty acids[90,91]. Many viral proteins also increased expression of genes involved in response to hypoxia (cluster C22).

Other characteristics that enable tumour growth include angiogenesis, invasion, and metastasis. Group III viral proteins induced increased transcription of genes in the VEGF pathway (cluster C23) and decreased transcription of genes involved in core functions of fibroblasts, including expression of collagen components, secreted growth factors and metalloproteases, and cell adhesion molecules such as integrins and protocadherins (clusters C13, C17, C18, C19; Supplementary Fig. 9). Such changes imply some amount of "reprogramming" induced by oncogenic viral proteins and again draw parallels with cancer pathogenesis. High-risk HPV proteins may achieve changes to cell-substrate adhesion and ECM production through transcriptional regulation of the growth factor EGR1. Other TFs with highly enriched binding sites for these pathway genes include FOSL2, which interacts with low-risk HPV E7s, and RUNX1. Still other viORFs may act through the transcription factors KLF7 and ARNTL, both of which are involved in insulin signalling[92,93].

Many of the transcriptional changes caused by viral proteins relate to the two enabling characteristics of cancer: genome instability and mutation, and tumour-promoting inflammation. Group III viral proteins caused decreased expression of genes involved in the response to DNA damage, including not only cell cycle arrest and apoptosis, but also autophagy via lysosome assembly (clusters C6 and C7) and acidification (cluster C3; 3.5-fold enriched for binding sites

of the autophagy implicated TF NRF2/NFE2L2[94] (*P* = 0.0007); and oxidative stress control through induction of G6PD and glutathione transferases GSTM4 and GSTM5 (cluster C15). We also observed marked upregulation of DNA nucleotide excision and double-stranded break repair (cluster C26 and C27: BRCA1, BRCA2, POL epsilon/delta, FANCB, XRCC2, BLM, RAD51) and robust activation of two "inflammasome" pathways known to be triggered by DNA damage: the type I interferon response via IRF activity (Cluster C24, 9.5-fold enriched, $P = 3 \times 10^{-10}$; Fig. 2b) and inflammatory responses (e.g. IL6, CCL2, IL1B) via NFκB activity (cluster C23, 4.7-fold enriched for NFκB sites, $P = 1.8 \times 10^{-5}$). Our network suggests that HPV E6s, EBV proteins, and polyomaviruses activate inflammatory pathways by regulating the expression of TFs like IRF1, NFkB, RELB, and MAFF. In contrast, HPV E7s perturb REL directly through a binary interaction. Tumour virus proteins also induce changes in the expression of genes involved in other signalling pathways, including PDGFR signaling, calcium signalling, and WNT signalling.

Pathway enrichment analysis of proteins identified through TAP-MS or Y2H was also instructive on established and novel pathways targeted by viruses. To identify host pathways targeted by viral proteins in our combined interactome dataset, we explored the enrichment of Gene Ontology (GO) terms among host target proteins (Supplementary Fig. 3). Enrichment for some of the GO terms was expected (*e.g.*, "mitotic cell cycle" for HPV, EBV and polyomavirus and "protein phosphatase 2A binding" for polyomavirus). Over-representation of "mitotic cell cycle" genes amongst HPV-associated proteins in our dataset arises not only from established associations of HPV E6 and E7 with the RB1 and TP53 tumour suppressors, but also from associations observed with 28 additional cell cycle proteins, including PCNA, MCM3 and CDKN1A ($P_{adj} < 0.01$, OR = 3.3) (Supplementary Table 17).

Enrichment analysis also revealed previously unrecognised pathways unique to specific viruses. EBV proteins collectively bind members of the procollagen-proline 4-hydroxylase family ($P_{adj} = 0.05$, OR = 389; Supplementary Fig. 3) including P4HA2, a hypoxia- and p53-responsive

protein, which inhibits angiogenesis and tumour growth in mice[95]. Adenoviral proteins bind multiple members of the BH-domain family (BAX, BIK, BAK1; P_{adj} = 0.04, OR = 159), which play a critical role in apoptosis[96]. The adenovirus E4ORF1 protein binds to a family of mRNA 5'UTR binding proteins (P_{adj} = 0.03, OR = 319), including the insulin-like growth factor binding protein 2 (IGFBP2), which has recently been implicated in angiogenesis and metastasis in breast cancer[97].

Supplementary References

30. Rual, J.F. *et al.*, Human ORFeome version 1.1: a platform for reverse proteomics. *Genome Res.* **14**, 2128-2135 (2004).
31. Sowa, M.E., Bennett, E.J., Gygi, S.P., & Harper, J.W., Defining the human deubiquitinating enzyme interaction landscape. *Cell* **138**, 389-403 (2009).
32. Cole, S.T. & Danos, O., Nucleotide sequence and comparative analysis of the human papillomavirus type 18 genome. Phylogeny of papillomaviruses and repeated structure of the E6 and E7 gene products. *J. Mol. Biol.* **193**, 599-608 (1987).
33. Dartmann, K., Schwarz, E., Gissmann, L., & zur Hausen, H., The nucleotide sequence and genome organization of human papilloma virus type 11. *Virology* **151**, 124-130 (1986).
34. Fuchs, P.G., Iftner, T., Weninger, J., & Pfister, H., Epidermodysplasia verruciformis-associated human papillomavirus 8: genomic sequence and comparative analysis. *J. Virol.* **58**, 626-634 (1986).
35. Zachow, K.R., Ostrow, R.S., & Faras, A.J., Nucleotide sequence and genome organization of human papillomavirus type 5. *Virology* **158**, 251-254 (1987).
36. Cole, S.T. & Streeck, R.E., Genome organization and nucleotide sequence of human papillomavirus type 33, which is associated with cervical cancer. *J. Virol.* **58**, 991-995 (1986).
37. Schwarz, E. *et al.*, DNA sequence and genome organization of genital human papillomavirus type 6b. *EMBO J.* **2**, 2341-2348 (1983).
38. Seedorf, K., Krammer, G., Durst, M., Suhai, S., & Rowekamp, W.G., Human papillomavirus type 16 DNA sequence. *Virology* **145**, 181-185 (1985).
39. Schwarz, E. *et al.*, Structure and transcription of human papillomavirus sequences in cervical carcinoma cells. *Nature* **314**, 111-114 (1985).
40. Sedman, S.A. *et al.*, The full-length E6 protein of human papillomavirus type 16 has transforming and trans-activating activities and cooperates with E7 to immortalize keratinocytes in culture. *J. Virol.* **65**, 4860-4866 (1991).
41. Liu, Y. *et al.*, Multiple functions of human papillomavirus type 16 E6 contribute to the immortalization of mammary epithelial cells. *J. Virol.* **73**, 7297-7307 (1999).
42. Schowalter, R.M., Pastrana, D.V., Pumphrey, K.A., Moyer, A.L., & Buck, C.B., Merkel cell polyomavirus and two previously unknown polyomaviruses are chronically shed from human skin. *Cell Host Microbe* **7**, 509-515 (2010).
43. van der Meijden, E. *et al.*, Discovery of a new human polyomavirus associated with trichodysplasia spinulosa in an immunocompromized patient. *PLoS Pathog.* **6**, e1001024 (2010).
44. Zalvide, J. & DeCaprio, J.A., Role of pRb-related proteins in simian virus 40 large-T-antigen-mediated transformation. *Mol. Cell. Biol.* **15**, 5800-5810 (1995).
45. Hahn, W.C. *et al.*, Enumeration of the simian virus 40 early region elements necessary for human cell transformation. *Mol. Cell. Biol.* **22**, 2111-2123 (2002).
46. Zhong, Q. *et al.*, Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.* **5**, 321 (2009).
47. James, P., Halladay, J., & Craig, E.A., Genomic libraries and a host strain designed for highly efficient two-hybrid selection in yeast. *Genetics* **144**, 1425-1436 (1996).
48. Walhout, A.J. & Vidal, M., A genetic strategy to eliminate self-activator baits prior to high-throughput yeast two-hybrid screens. *Genome Res.* **9**, 1128-1134 (1999).
49. Rual, J.F. *et al.*, Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-1178 (2005).
50. Venkatesan, K. *et al.*, An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83-90 (2009).

51. Litovchick, L., Florens, L.A., Swanson, S.K., Washburn, M.P., & DeCaprio, J.A., DYRK1A protein kinase promotes quiescence and senescence through DREAM complex assembly. *Genes Dev.* **25**, 801-813 (2011).
52. Spangle, J.M. & Munger, K., The human papillomavirus type 16 E6 oncoprotein activates mTORC1 signaling and increases protein synthesis. *J. Virol.* **84**, 9398-9407 (2010).
53. Baker, S.J., Markowitz, S., Fearon, E.R., Willson, J.K., & Vogelstein, B., Suppression of human colorectal carcinoma cell growth by wild-type p53. *Science* **249**, 912-915 (1990).
54. Litovchick, L. *et al.*, Evolutionarily conserved multisubunit RBL2/p130 and E2F4 protein complex represses human cell cycle-dependent genes in quiescence. *Mol. Cell* **26**, 539-551 (2007).
55. Livak, K.J. & Schmittgen, T.D., Analysis of relative gene expression data using real-time quantitative PCR and the 2(-ΔΔCT) Method. *Methods* **25**, 402-408 (2001).
56. Stewart, S.A. *et al.*, Lentivirus-delivered stable gene silencing by RNAi in primary cells. *RNA* **9**, 493-501 (2003).
57. Ramsby, M.L. & Makowski, G.S., Differential detergent fractionation of eukaryotic cells. Analysis by two-dimensional gel electrophoresis. *Methods Mol. Biol.* **112**, 53-66 (1999).
58. Nakatani, Y. & Ogryzko, V., Immunoaffinity purification of mammalian protein complexes. *Methods Enzymol.* **370**, 430-444 (2003).
59. Ficarro, S.B. *et al.*, Improved electrospray ionization efficiency compensates for diminished chromatographic resolution and enables proteomics analysis of tyrosine signaling in embryonic stem cells. *Anal. Chem.* **81**, 3440-3447 (2009).
60. Parikh, J.R. *et al.*, multiplierz: an extensible API based desktop environment for proteomics data analysis. *BMC Bioinformatics* **10**, 364 (2009).
61. Webber, J.T., Askenazi, M., & Marto, J.A., mzResults: an interactive viewer for interrogation and distribution of proteomics results. *Mol. Cell. Proteomics* **10**, M110 003970 (2011).
62. Askenazi, M., Marto, J.A., & Linial, M., The complete peptide dictionary--a meta-proteomics resource. *Proteomics* **10**, 4306-4310 (2010).
63. Askenazi, M., Li, S., Singh, S., & Marto, J.A., Pathway Palette: a rich internet application for peptide-, protein- and network-oriented analysis of MS data. *Proteomics* **10**, 1880-1885 (2010).
64. Chatr-aryamontri, A. *et al.*, VirusMINT: a viral protein interaction database. *Nucleic Acids Res.* **37**, D669-673 (2009).
65. Berriz, G.F., Beaver, J.E., Cenik, C., Tasan, M., & Roth, F.P., Next generation software for functional trend analysis. *Bioinformatics* **25**, 3043-3044 (2009).
66. Dai, M. *et al.*, Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33**, e175 (2005).
67. Johnson, W.E., Li, C., & Rabinovic, A., Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-127 (2007).
68. Cheung, M.S., Down, T.A., Latorre, I., & Ahringer, J., Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Res.* **39**, e103 (2011).
69. Benjamini, Y. & Hochberg, Y., Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B (Methodological)* **57**, 289-300 (1995).
70. Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., & Ideker, T., Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431-432 (2011).
71. Pique-Regi, R. *et al.*, Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21**, 447-455 (2011).
72. Palomero, T. *et al.*, NOTCH1 directly regulates *c-MYC* and activates a feed-forward-loop transcriptional network promoting leukemic cell growth. *Proc. Natl. Acad. Sci. USA* **103**, 18261-18266 (2006).

73. Raychaudhuri, S. *et al.*, Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5**, e1000534 (2009).
74. Myers, S., Bottolo, L., Freeman, C., McVean, G., & Donnelly, P., A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321-324 (2005).
75. Starr, T.K. *et al.*, A Sleeping Beauty transposon-mediated screen identifies murine susceptibility genes for adenomatous polyposis coli (Apc)-dependent intestinal tumorigenesis. *Proc. Natl. Acad. Sci. USA* **108**, 5765-5770 (2011).
76. March, H.N. *et al.*, Insertional mutagenesis identifies multiple networks of cooperating genes driving intestinal tumorigenesis. *Nat. Genet.* **43**, 1202-1209 (2011).
77. Starr, T.K. *et al.*, A transposon-based genetic screen in mice identifies genes altered in colorectal cancer. *Science* **323**, 1747-1750 (2009).
78. Mann, K.M. *et al.*, Sleeping Beauty mutagenesis reveals cooperating mutations and pathways in pancreatic adenocarcinoma. *Proc. Natl. Acad. Sci. USA* **109**, 5934-5941 (2012).
79. Bergerson, R.J. *et al.*, An insertional mutagenesis screen identifies genes that cooperate with Mll-AF9 in a murine leukemogenesis model. *Blood* **119**, 4512-4523 (2012).
80. Cancer Genome Atlas Research Network, Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615 (2011).
81. Agrawal, N. *et al.*, Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in *NOTCH1*. *Science* **333**, 1154-1157 (2011).
82. Chapman, M.A. *et al.*, Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467-472 (2011).
83. Lee, W. *et al.*, The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473-477 (2010).
84. Parsons, D.W. *et al.*, An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807-1812 (2008).
85. Parsons, D.W. *et al.*, The genetic landscape of the childhood cancer medulloblastoma. *Science* **331**, 435-439 (2011).
86. Pleasance, E.D. *et al.*, A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184-190 (2010).
87. Puente, X.S. *et al.*, Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475**, 101-105 (2011).
88. Stransky, N. *et al.*, The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157-1160 (2011).
89. Kuehnle, K. *et al.*, Prosurvival effect of DHCR24/Seladin-1 in acute and chronic responses to oxidative stress. *Mol. Cell. Biol.* **28**, 539-550 (2008).
90. Scaglia, N., Caviglia, J.M., & Igal, R.A., High stearoyl-CoA desaturase protein and activity levels in simian virus 40 transformed-human lung fibroblasts. *Biochim. Biophys. Acta* **1687**, 141-151 (2005).
91. Scaglia, N. & Igal, R.A., Stearoyl-CoA desaturase is involved in the control of proliferation, anchorage-independent growth, and survival in human transformed cells. *J. Biol. Chem.* **280**, 25339-25349 (2005).
92. Kawamura, Y., Tanaka, Y., Kawamori, R., & Maeda, S., Overexpression of Kruppel-like factor 7 regulates adipocytokine gene expressions in human adipocytes and inhibits glucose-induced insulin secretion in pancreatic beta-cell line. *Mol. Endocrinol.* **20**, 844-856 (2006).
93. Marcheva, B. *et al.*, Disruption of the clock components CLOCK and BMAL1 leads to hypoinsulinaemia and diabetes. *Nature* **466**, 627-631 (2010).

94. Rao, V.A. *et al.*, The antioxidant transcription factor Nrf2 negatively regulates autophagy and growth arrest induced by the anticancer redox agent mitoquinone. *J. Biol. Chem.* **285**, 34447-34459 (2010).
95. Teodoro, J.G., Parker, A.E., Zhu, X., & Green, M.R., p53-mediated inhibition of angiogenesis through up-regulation of a collagen prolyl hydroxylase. *Science* **313**, 968-971 (2006).
96. Kelekar, A. & Thompson, C.B., Bcl-2-family proteins: the role of the BH3 domain in apoptosis. *Trends Cell Biol.* **8**, 324-330 (1998).
97. Png, K.J., Halberg, N., Yoshida, M., & Tavazoie, S.F., A microRNA regulon that mediates endothelial recruitment and metastasis by cancer cells. *Nature* **481,** 190-194 (2011).